TECHNISCHE UNIVERSITÄT MÜNCHEN
INSTITUT FÜR INFORMATIK

Lehrstuhl für Angewandte Informatik / Kooperative Systeme

Mobile Social Situation Detection

Dipl. Inf. Univ. Alexander Friedrich Wilhelm Lehmann

Vollständiger Abdruck der von der Fakultät für Informatik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktor der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender:	Prof. Dr. Hans-Joachim Bungartz
Prüfer der Dissertation:	1. Prof. Dr. Johann Schlichter
	2. Prof. Dr. Helmut Seidl
	3. Priv.-Doz. Dr. Georg Groh

Die Dissertation wurde am 3.11.2015 bei der Technischen Universität München eingereicht und durch die Fakultät für Informatik am 25.2.2016 angenommen.

For Friedwart.

ABSTRACT

Research and applications in the field of Social Signal Processing have successfully targeted audio- and video-based techniques for the extraction and interpretation of behavioural cues. Corresponding research typically aims at specific and rather narrow scenarios, often limited by dependencies on external infrastructure. This thesis investigates the detection of social situations based on mobile devices and numerous physical or logical sensors, preferably without any such dependencies. More specifically, it is shown how probability models based on variables of human interaction geometry lead to reliable results for the detection and classification of binary social interaction, from which logical deduction and sensor fusion lead to the determination of n-ary social situations. The input data of possible real-life applications are mined from mobile sensors, resulting in wider applicability. A new research dataset is aquired in a laboratory experiment by precise detection of interaction geometry through a commercial infrared tracking system, bypassing the difficulties involved in mobile sensing and allowing for more fine-grained error analysis for the real-life application case. Potential influences of personal profile parameter and latent variables such as gender and group size onto the model are investigated using an additional new dataset. The applicability of the proposed model in mobile scenarios is evaluated based on two new mobile systems for measurements of interaction geometry. Interaction geometry is however mainly well-suited for the analysis of static situations. The second part of this thesis hence demonstrates the ability to recognize dynamic situations by means of dual co-activity detection based on the similarity of multivariate data streams from mobile agents, i. e. the detection of co-located and -timed activities of the same type, consequently serving as indicators for the presence of mutual social situations. Contrary to related research in the field of Activity Recognition, this new approach does not explicitly classify the actual activities, as the detection of arbitrary co-activities out of the unlimited spectrum of potential activities would otherwise be limited by application- or research-specific heuristics.

ZUSAMMENFASSUNG

Forschung und Anwendung im Bereich des Social Signal Processings haben erfolgreich audio- und videobasierte Verfahren zur Extraktion und Interpretation von Behavioural Cues entwickelt. Die Forschung beschäftigt sich diesbezüglich üblicherweise mit spezifischen und eher begrenzten Szenarien, oftmals limitiert durch Abhängigkeiten von externer Infrastruktur. Die vorliegende Arbeit untersucht die Erkennung von sozialen Situationen basierend auf Mobilgeräten und zahlreichen physikalischen und logischen Sensoren, vorzugsweise ohne vorgenannte Abhängigkeiten. Es wird gezeigt, wie Wahrscheinlichkeitsmodelle, basierend auf Variablen menschlicher Interaktionsgeometrie, zu verlässlichen Ergebnissen hinsichtlich der Detektierung und Klassifikation von binärer sozialer Interaktion führen. Logische Deduktion sowie die Vereinigung mehrerer Sensoren führen dann zur Bestimmung von n-ären sozialen Situationen. Die Eingabedaten möglicher echter Anwendungen werden aus mobilen Sensoren gewonnen, wodurch das Einsatzgebiet erweitert wird. Ein neuer Datensatz zur Forschung wird in einem Laborexperiment durch die präzise Messung von Interaktionsgeometrie mithilfe eines kommerziellen Infrarot-Trackingsystems erzeugt, um Schwierigkeiten und Ungenauigkeiten im Rahmen mobiler Erfassung dieser Daten zu umgehen und eine feingranulare Fehleranalyse für den Einsatz in echten Anwendungen zu ermöglichen. Mögliche Einflüsse auf das Modell durch persönliche Profilparameter und latente Variablen, wie beispielsweise Geschlecht und Gruppengröße, werden anhand eines weiteren neuen Datensatzes untersucht. Die Anwendbarkeit des Modells in mobilen Szenarien wird anhand zweier neuer mobiler Systeme zur Messung von Interaktionsgeometrie evaluiert. Nichtsdestotrotz ist Interaktionsgeometrie hauptsächlich zur Analyse statischer Situationen geeignet. Der zweite Teil dieser Arbeit zeigt daher, wie sich dynamische Situationen auf Basis dualer Co-Aktivitäten erkennen lassen, basierend auf Ähnlichkeitsmaßen zwischen den multivariaten Datenströmen zweier mobiler Agenten. Hierbei dient die Erkennung von Co-Aktivitäten als orts- und zeitgleiche Aktivitäten desselben Typs als Indikator für die Existenz sozialer Situationen. Im Gegensatz zum Gebiet der Activity Recognition werden in diesem neuen Verfahren die Aktivitätstypen aber nicht explizit klassifiziert, um die Erkennung beliebiger Co-Aktivitäten aus einem unbegrenzten Spektrum möglicher Aktivitäten nicht durch applikations- oder forschungsspezifische Heuristiken zu limitieren.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

AICc	Akaike Information Criterion corrected	65
AIC	Akaike Information Criterion	65
API	Application Programming Interface	127
AR	Activity Recognition	2
BIC	Bayesian Information Criterion	64
BMA	Belief Mass Assignment	170
BN	Bayesian Network	168
CV	Computer Vision	141
DBMA	Dirichlet Belief Mass Assignment	173
DBN	Dynamic Bayesian Network	191
DCM	Direction Cosine Matrix	22
DCT	Discrete Cosine Transform	205
DR	Dead Reckoning	132
DST	Dempster-Shafer theory	170
EKF	Extended Kalman Filter	134
EM	Expectation Maximization	18
ENU	East-North-Up	128
FACS	Facial Action Coding System	9
FAR	False Alarm Rate	220
FAS	Face Address System	13
FFS	F-Formation System	12
FFT	Fast Fourier Transform	190
GMM	Gaussian Mixture Model	50
GPS	Global Positioning System	8
GTM	Generative Topographic Mapping	194
HARP	Human Activity Recognition Problem	196
HCI	Human Computer Interface	17
HMM	Hidden Markov Model	49
INS	Inertial Navigation System	133
IN	Inertial Navigation	133
KF	Kalman Filter	129
KNN	K Nearest Neighbour	184
LVM	Latent Variable Model	50
MDR	Missed Detection Rate	220
MEMS	Micro-Electro-Mechanical System	129
MFCC	Mel Frequency Cepstral Coefficient	180
MSN	Mobile Social Networking	167

NCA	Neighbourhood Component Analysis	194
NMF	Non-Negative Matrix Factorization	194
O-space	Object Space	12
P-space	Personal Space	12
PCA	Principal Component Analysis	61
PF	Particle Filter	130
R-space	Rear Space	12
RMS	Root Mean Squared	130
SL	Subjective Logic	XVII
SMEFO	Socially Motivated Estimate of Focus Orientation	18
SN	Social Network	180
SSP	Social Signal Processing	1
STFT	Short-Term Fourier Transform	148
SVD	Singular Value Decomposition	54
SVM	Support Vector Machine	48
SW-GMM	Semi-Wrapped Gaussian Mixture Model	57
TA	Transactional Segment	12
UDID	Unique Device Identifier	199
W-GMM	Wrapped Gaussian Mixture Model	57
i.i.d.	independent and identically distributed	51
w.l.o.g.	without loss of generality	61

1 INTRODUCTION / MOTIVATION

The technological advancements of computers, portable and mobile devices, tablets and other gadgets, as well as of course the developments in the related disciplines in science and engineering, have led to a state where computing, networking, monitoring and a vast amount of services seem to be omni-present throughout society. Next to typical candidates such as smartphones, mobile devices are to be found literally anywhere, even in clothing, known as wearables, and recently also smartwatches [182, 225, 207, 230, 146]. The corresponding field is known as *ubiquitous* or *pervasive computing* [281], where researchers and engineers alike strive for the seamless integration of technology into various aspects of daily life and human routine, thereby creating a transparent link between virtual and real life [277, 295]. Out of all of the aforementioned devices, mobile phones have had the highest adoption rate during the past two decades [115]. In 2013, already more than 90% of the German and 91% of the American population owned one or more mobile phones [322, 258]. The omni-presence of mobile devices has notably influenced the way that people interact with each other. Nowadays, social interaction is no longer confined to co-located activities of the participating persons, but extends to deferred location just as well as to deferred time. Arminen et al. refer to this development as the "reformation of social actions in mobile space-time" [14].

Computers have advanced from mere data processors to interaction partners of humans. Resources and channels like the web, email, messaging services, mobile applications, and social networking platforms, to name only a few, have elevated computers to a "privileged interaction medium for social exchange" [278]. As possibly even proactive interactants, computers should therefore learn to understand and synthesize social signals in order to provide better communication, interaction and contextual services [120].

Likewise, based on the insight that people communicate using a "subtle combination of gesture, facial expression, body language, and vocal prosody in conjunction with spoken words" [230], Pentland coined the term "perceptually aware" for machines and environments that would be able to understand and generate such communicative elements to obtain improved human-computer interfaces [230]. This is also known as *socially aware computing* [231, 25, 195]. Two of its major aspects are the research of, and applications for, the constantly increasing number of available uni- and multimodal sensors on mobile platforms, in particular mobile phones. According to Lane et al., this may eventually revolutionize economical sectors such as "business, healthcare, social networks, environmental monitoring, and transportation" [181], since "mobile phone sensing systems will ultimately provide both micro- and macroscropic views of cities, communities, and individuals, and help improve how society functions as a whole" [181]. Among others, these are just some of the key motivators for research disciplines such as Social Signal Processing (SSP) and

Activity Recognition (AR), which make substantial efforts to fusion the findings of social sciences with those of pervasive computing, sensor networks, data mining, machine learning and algorithmic modeling.

1.1 SOCIAL SIGNAL PROCESSING

Generally speaking, SSP corresponds to the analysis of *non-verbal* human-human and human-computer interaction, for which computers may be considered as *social actors* [334]. The primary goal of SSP is to help computers to develop the abilities to recognize and understand human social signals [333]. This can also involve the synthesis of social signals during phases of active acting and back-channeling.

At its heart, SSP is based on the consideration that social intelligence is a key factor for success [334]. Being able to understand, express and manage social signals would help interactants to “get along well with others while winning their cooperation” [333]. In other words, aiding humans and computers alike in their understanding of social signals may further allow them to *exploit* this knowledge to become more effective when dealing with social interactions, e. g. when assuming a dominant role in a social group or the working environment, being more successful at job interviews or business transactions, etc. [232, 334]. As such, it may help to find the right margin between acting appropriately or inappropriately during social interactions, [12] in [334].

Existing and possible applications of SSP include the analysis of the spread of diseases or the flow of information based on social network analysis [230, 84, 120], urban planning and traffic forecasting [117], social lifelogging [124, 187], sharing content through social participation [106], multimedia indexing [345, 303], analysis of human privacy bounds [70], crowd sensing and crowd sourcing [105], crowd motion analysis [151, 151, 217, 355], monitoring devices for law-enforcement officers and firefighters [98, 49], or increasing the efficiency of call-centers by prematurely ending a call that has no prospect of ending in the acquise of a new customer [232]. SSP could furthermore act as a service layer for contextual social networking [120], privacy management, reachability management, controlling the flow of information, e. g. when addressing social groups without specific interest in individual members of those groups.

1.1.1 *Social Signals*

Pentland was the first to establish the notion of *social signals*, described in [232]:

> “Social signaling is what you perceive when observing a conversation in an unfamiliar language and yet find that you can still *see* someone taking charge of a conversation or establishing a friendly interaction.”

This example was later rephrased by Vinciarelli in terms of being able to “capture the social landscape” despite a lack of verbal understanding [334]. This implies that social signals

provide an independent, quantifiable channel of communication [232] which constitutes as much as 90% of human communication, albeit varying with context [333].
A more formal definition is given by Poggi and D'Errico in [243, 244], and repeated by Vinciarelli in [336]:

> "A social signal is a communicative or informative signal that, either directly or indirectly, provides information about social facts, namely, social interactions, social emotions, social attitudes, or social relations."

For this, a *social action* is defined as any event performed by an agent A in relation to another agent B, where A considers B as a self-regulated agent with subjective goals [336]. A *social interaction* is consequently defined as a social action that is performed by A while B is actually or virtually present, [333] in [336]. According to Salah et al., social signals extend beyond the real world, e. g. in contexts such as micro-blogging or connection formations over social networking platforms [278].
Note that the given definition of social signals distinguishes between *informative* and *communicative* signals. The notion is that communicative signals are emitted consciously along with verbal expressions in order to provide the addressee with contextual information for the subsequent interpretation of the received message, as for example the choice of tone when saying something ironical. On the other hand, informative signals may be emitted consciously or unconsciously, yet always carry information that is received and interpreted by the addressee [336]. A particularly interesting point about unconscious signals is their tendency to convey *honest* information [233, 336]. As unconscious signals, they are also not explicitly controllable [334]. According to Salah et al., humans are "evoluationary bound to produce social signals", even when they are alone [278]. This is further sustained by Knapp et al. who state that people tend to use gestures even when they are alone or on the phone, i. e. when no recepient is actually present [173]. Vinciarelli et al. further mention another category of signals, namely those that are actually not emitted per se. Among other examples, this is the case for mimicry, i. e. when a person A assumes the posture (gestures, etc.) of another person B, from which a third party C could tell that either A or B are mimicking the respective other, provided that C can observe *both* persons at the same time. This would not be possible for anyone observing only either one of A *or* B [336]. Signals of the latter category can therefore be considered both communicative and informative.

1.1.2 *Behavioural Cues*

Social signals correspond to a temporal superposition of non-verbal *behavioural cues*, typically lasting for a short time in the range of milliseconds to minutes [336, 333, 278, 232]. As such, behavioural cues can be regarded as atomic units even though their complexity may vary. Well acknowledged behavioural cues usually fall into either one of the following categories, for each of which a detailed overview is given in [333]:

- Physical appearance, e. g. height, attractiveness, body shape

- Gesture and posture, e. g. hand gestures, posture, walking
- Face and eyes behaviour, e. g. facial expressions, gaze, focus of attention
- Vocal behaviour, e. g. prosody, turn taking, vocal outbursts, silence
- Space environment, e. g. seating arrangement, distance, orientation

Humans are particularly effective in recognizing these behavioural cues (and numerous more). Extending the notion to social signals or social factors in general, parts of the human brain, so-called mirror neurons, are in fact specialized in the recognition and processing of such factors as well as in the awareness of other social interactants [271]. Behavioural cues hence produce "social awareness, i. e. a spontaneous understanding of social situations that does not require attention or reasoning", [177] in [333].

In the context of SSP, the principle advantage of behavioural cues over (the more complex) social signals is that they can be automatically detected and recognized by means of rather simple sensors such as cameras and microphones [334, 333, 232]. Furthermore, behavioural cues can be captured without precise knowledge about the context in which they appear. According to Kendon and Scheflen, gaze, focus, posture changes, etc. have no intrinsic meaning at all [166, 282]. In turn, this implies that the actual understanding of social signals is inevitably bound to *context-sensitive* interpretation. This is corroborated by Pentland who postulates that "useful systems must be able to adjust for individual differences, become more sensitive to task and environmental constraints, and be able to relate face and hand gestures to the semantics of the human-machine or human-human interaction" [230]. They are not specifically linked to linguistic structures or affective states [232].

1.1.3 *Latent Information in Non-Verbal Behaviour*

It was mentioned before that social signals provide an independent, quantifiable channel of information. Unconscious, potentially honest, non-verbal behaviour furthermore constitutes a continuous source of insight into personal feelings, mental state and personality [269]. Social signals therefore convey information about a subject's *inner state.*

Mohammadi et al., for example, investigated the automatic analysis of personality traits based on individual samples of speech [215]. For this, they compared the algorithmic results with those of human judgements on a relative large corpus consisting of 640 samples. To ensure that the interpretation would be restricted to non-verbal communication, samples were chosen such that the human experimenters would not understand the respectively spoken languages. Personality traits were then expressed on a scale using the Five Factor Model [203], according to which each traits falls into one of the categories extraversion, agreeableness, conscientousness, neuroticism, and openness (see section 2.4.1 for further descriptions of the model). Although the performance of the demonstrated system was mixed, it still shows that automatic assessment of at least some individual personality traits is feasible. Other works in the field [238] achieved as much as 94% accuracy on a much larger corpus with respect to extraversion and locus-of-control, the latter of which quantifies if personal behaviour is assumed to be dependent on a subject's own actions or on external factors [107].

However, other than the analysis of individual persons, social signals also tell us something about the nature and quality of *interpersonal relationships* during social interactions [333]. Examples are given by the way that people are oriented towards each other while they are talking, whether they remain silent, how they position themselves, what intonation they choose, their mutual turn-taking patterns and whether there are overlapping segments of speech, their visual focusses, etc. It follows that by continuously detecting and analyzing social signals, information about individuals as well as their relationships can be automatically inferred on a much more fine-granular scale, both in terms of quality and timeliness, than, for example, based on long-term and overly simplified (yes/no) information such as friendship relations on social networking platforms. Another advantage is that the information contained in social signals is *implicit*. In contrast, explicit assessments of relationships by humans are prone to misjudgement. Such misjudgements may be conscious or unconscious, e. g. due to efforts to avoid the violation of social conventions and therefore personal embarassment. People may also have difficulties when being asked to explicitly assess their relations with others (see section 2.4.1). The latter may be due to subjectively varying scale, or due to the fact that people have a different impression of the quality of their relation. For example, a person A might deem another person B as a close friend, whereas B might see A as acquaintance. Even in the case where A and B would exhibit mutual agreement on being friends, their personal understandings of "friend" might differ. To the contrary, automatic recognition and interpretation of social signals, provided that (possibly domain-specific) means of standardization exist, based e. g. on well-acknowledged findings from fields such as sociopsychology and sociology, will most likely yield much clearer and hopefully universally applicable results, together with a significant increase in precision.

1.1.4 *Main Objectives of Social Signal Processing*

SSP augments the classical approach of asking about the Where, What, When, and Who [226, 135, 191] with questions about the How and Why [336], thereby enriching apparent perception by context-sensitive social aspects such as communicative intention, affect, and cognitive state. In order to do so, SSP is concerned with the following core questions [243, 336, 6]:

- How to algorithmically detect and recognize behavioural cues based on uni- and multimodal sensors such as cameras, microphones, or accelerometers?
- How to algorithmically infer social signals and attributes from these behavioural cues?
- How to synthesize social signals in an effort to improve socially aware computing and human-computer interfaces?

Although not explicitly listed, this enumeration naturally implies the aspect of modeling. Therefore the three main objectives of SSP can be stated as the modeling, analysis and synthesis of social signals. SSP has evolved from initially mostly speech- and computer-vision based systems, which were able to "detect, track, and identify people and more generally, to interpret human behaviour" [230], to more complex systems based on often sophisticated models as well as the integration and fusion of arrays of multiple physical and logical sensors, hence "potentially permitting computer and communication systems to support social and organizational roles instead of viewing the individual as an isolated entity" [232]. As such, SSP strives for the development of tools and models which accurately capture and/or predict human behaviour. Hitherto research has shown that corresponding tools may sometimes even exceed expert human capabilities [232]. It is therefore expected that SSP will empower research with the ability to achieve results with much higher quality and in shorter time [334, 232, 336, 278]. Yet SSP should not be seen as an orthogonal means that would eventually replace human experts in their fields. However, contrary – or in addition – to human experts, SSP will likely yield more objectivity during interpretation of what was observed. It does furthermore allow for observations on a much larger quantitative scale, to be used e. g. in aiding researchers of fields other than in natural sciences.

1.1.5 *Predictability of Human Behaviour*

An important question in this matter is whether human behaviour can be modeled or predicted accurately. Generally speaking, it is certainly true that human behaviour is, in principle, innumerable in terms of versatility. On the other hand, some portions of human behaviour are more likely to contribute to social interaction and social signaling than others. Also, it is assumed that some behavioural patterns are likely to occur more often than others. In other words, while there is potential for random patterns in human behaviour, there are also identifiable routines [82]. Without doubt, interpretation and validity of the latter is a matter of the problem-specific application domains in SSP.

A number of studies however show that modeling and prediction are indeed possible. Song et al., for example, studied how much traces of human mobility were predictable by measuring the entropy of the location trajectories of $\sim 50,000$ users, for which they report a 93% potential for correct predictions. Perhaps surprisingly, they found that none of the subjects was predictable with less than 80%, even though parameters such as age, gender, population density or travel habits varied widely between subjects [310]. In general, subjects would spend 45% of their time at their primary location, and between 60% to 80% at their second to tenth most visited locations [310]. On a sidenote, studies like the former also sustain the principle feasibility of universally applicable models both in terms of gathering enormous datasets as well as exploiting those data. Data on human mobility, for instance, are vastly collected by mobile phone carriers across large segments of the population [117].

Other, particularly well acknowledged studies were performed by Eagle and Pentland, based on their *Reality Mining* dataset [82, 83]. The dataset consists of the recordings of 100 mobile phones over the course of nine months, accounting for $\sim 450,000$ hours of information about the users. Notably, merely 15% of data are missing or uncontinuous due to battery depletion or the fact that users consciouscly turned off their phones. Informations recorded include the devices' call logs, application usage, key presses, bluetooth devices in proximity and cell towers. While Eagle and Pentland could again estimate location with great accuracy, they were furthermore able to infer the social network and quantify the subjects' mutual relationships with over 90% accuracy, differentiating between workspace colleagues, outside friends and people within a circle of friends [82]. It is particularly worth noting that in order to infer friendship from daily proximity networks, the *context* of each mutual encounter had to be taken into account, more precisely location and time of their proximity measures.
In a subsequent study, Eagle and Pentland showed the prevalence of routine in the daily lives of their subjects [83]. For this, they determined the principal components of the recorded data on varying time scales. While these components correspond to the so-called *eigenbehaviour* of an individual, this concept can easily be extended to groups, eventually providing a similarity measure for individuals as well as for groups. Eagle and Pentland report reconstruction accuracies of more about 80% using only a single eigenvector, already more than 90% for five eigenvectors, and eventually close to 100% for as little as 15 eigenvectors. All the same, combining the eigenvectors of the individuals of certain groups, which would consequently span the so-called *behaviour space*, they further determined that the behaviour of "first year students" was the most predictive, whereas that of "business school students" was the least. The latter analysis contributes to Eagle and Pentland's notion of lives' entropy: "People who live entropic lives tend to be more variable and harder to predict, while low-entropy lives are characterized by strong patterns across all time scales" [82]. Another notable result of their study is the fact that the identification of a mere 50% of individual's recording in terms of his or her eigenbehaviour would be sufficient to predict the remaining 50% with 79% accuracy [83].

Apart from the prediction of long-term behaviour, another interesting aspect is the analysis of short-term behavioural cues. This can, for instance, be used to get insight into the inner state of a conversational partner, and e. g. exploit that knowledge towards a successful aquise of a new customer at a call-center. A corresponding experiment monitored 70 calls and subsequently analyzed social signals such as engagement, mirroring, activity and stress, based solely on the tone of voice [232]. These signals were then used to predict the outcome of a proposed deal, yielding an overall accuracy of about 87% for successful predictions of the outcome of a corresponding call.

1.1.6 *What is Recorded?*

An increasing number of datasets exist for research in SSP, only some of which are available to the public, for example [66, 82, 121]. Overviews of further datasets can be found e. g. in [107, 333]. In addition to datasets, an even more increasing number of frameworks aim at SSP tasks [11, 237, 47, 15, 211, 280, 252]. The common denominator of these frameworks is their attempt to provide the user with more or less easy access to physical and logical mobile phone sensors such as Global Positioning System (GPS) receivers, compasses, accelerometers, gyroscopes, thermometers, barometers, cell towers, Bluetooth, microphones, cameras, WLAN, call logs, SMS logs, applications, contacts, history, battery state, near field communication, etc. Some of these frameworks also manage encryption, or remote configuration and survey systems, the latter of which may be useful for online annotation of the recorded data by the monitored subjects. Yet other frameworks attempt to provide solutions for energy efficient handling of sensors as well as minimizing the effect of monitoring on device performance. Last but not least, some of the frameworks extend beyond those services and provide basic functionality for the automatic detection and/or classification of events.

The exemplary enumeration of sensors available in modern mobile devices sustains the extent of possible applications in SSP research. The expanding number of embedded sensors is also considered on the key drivers of mobile phone applications [181]. Yet in spite of their sheer numbers, the selection and interaction of sensors is just as important for the respective problem domain. The kinds of social signals and behavioural cues that were recorded will consequently constitute an upper bound of what can be learned from the data [335], and in general the acquisition of well-suited and large enough datasets is a tedious and time consuming process. The design of experiments and algorithmical models should furthermore take into account that social signals are considered to be “intrinsically ambigious and the best way to deal with such problem is to use multiple behavioural cues extracted from multiple modalities” [333]. Models may also benefit from learning the “texture” of certain social signals instead of trying to understand the actual signals themselves [232]. Depending on the application it may be sufficient to learn the correlation between social signals and the investigated entities, such as e. g. the outcome of acquiring a new customer in the call-center example.

The majority of mobile phone sensors can be classified as either inertial, positioning or ambient [144], but sensors may also fall into more than one category. Bluetooth, for instance, may be used for estimating distance or indoor localization scenarios [52, 46, 24], yet also for inferring social networks or routine behaviour based on past encounters [82, 83], or other interesting approaches like analyzing the sets of people who frequently encounter or see each other without mutual awareness or even knowing each other [228].

Similarly, microphones are extremely versatile in regard of the detection of ambient noise, silence, verbal and/or non-verbal expressions, turn-taking patterns, prosody, energy, vocal outburts, etc. Groh et al., for example, successfully analyzed turn-taking patterns in order to infer social interaction [126].

Last but not least, one of the most frequently deployed sensors are cameras. Still images or continuous recordings can convey an enormous amount of information, and research in computer vision has long since demonstrated the ability to detect, recognize and possibly track gestures, postures, faces, eyes, ears, mouths, extremities, space and environment, seating arrangement or objects in general [191, 329, 214, 337, 136]. Among the former, the face is particularly important as it hosts the greatest part of our most important sensoric organs. Gaze, for instance, can tell a lot about the quality of social interaction [165], the eyes may help to distinguish real smiles from fake ones [75], and together with other parameters, the face can be analyzed to detect deception or lies [99].

It is interesting to realize that facial expressions can be almost exhaustively described in terms of the so-called Facial Action Coding System (FACS) [86], comprised of a surprisingly small number of *action units* and *action descriptors*, based on individual or groups of muscles and their corresponding movements, respectively. Likewise, it is also possible to describe the signals of sign languages with rather basic sets of parameters [336]. Next to capturing such kinds of behavioural cues, developing a higher-level concept for the description of abstract parameters such as their respective amplitude, fluidity, power, acceleration and repetition could yield a useful grammer to lift those behavioural cues into the context of social signals [336].

1.1.6.1 *Obtrusive Sensing*

Generally speaking, SSP is about making *implicit* facts *explicit*, for which examples were given such as a person's inner state or their relationship towards others. The corresponding social signals are mostly unconscious, which implies that explicit interaction between a device and its user should be avoided, as that may lead to erroneous and biased observations, as well as alter the users' behaviour when they are aware of the fact that they are being recorded. For example, traditional vision-based approaches, as opposed to using e. g. inertial sensors, can be considered intrusive and disruptive [16].

Whenever active support of the sensing process is required by the user, this is known as *participatory sensing*, whereas mere passive involvement would be known as *opportunistic sensing* [181, 144]. It follows that in scenarios where emphasis lies on unconscious and/or honest behaviour, such as e. g. in the analysis of social interaction geometry on small spatio-temporal scales, opportunistic sensing would provide the appropriate means. Extending to larger settings, opportunistic sensing can also be considered "particularly useful for community sensing, where per user benefit may be hard to quantify and only accrue over a long time" [181]. Participatory sensing, on the other hand, may help to increase the acceptance level of sensing applications with respect to privacy [144]. In that sense, people should always be aware of the fact that they are being recorded and that they are possibly sharing data. This would have the advantage that people could decide what type and amount of information they are willing to share, provided of course that they are basically capable of realizing the consequences and implications with respect to subsequent analysis of their personal data. Some researchers therefore propose that especially raw sensor data should generally not be pushed to the cloud because of any associated, non-foreseeable privacy issues [181]. At last, people should in principle constantly be able to enact their proprietary rights, possibly even including an opportunity for posterior erasure of the data. Surprisingly though, the enormous developments of social networking platforms and the exponential availability of data however lead to the presumption that people indeed have an intrinsic motivation and willingness to share nearly all kinds of personal and non-personal information, accepting the risk that it may be exploited both legally and illegaly.

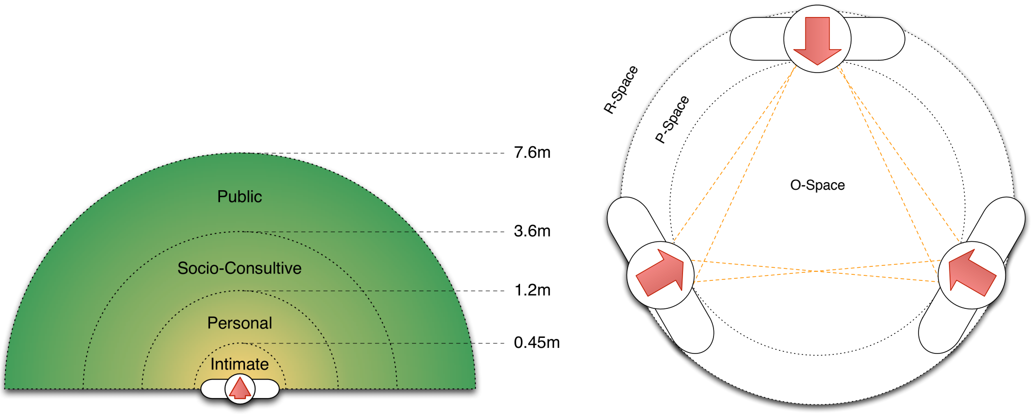
1.2 INFERRING SOCIAL INTERACTION FROM SPATIO-ORIENTATIONAL ARRANGEMENTS

Social relationships can be regarded as a function of social interaction [166]. This may be based on quantitative and/or qualitative measures such as the presence or absence of interaction, relative frequency, respective duration, interpersonal distances during interaction, or temporal behavioural patterns such as trajectories or the ratio of individual versus the sum of interactions. Subsuming, social relationships can be characterized by investigating how they were built and sustained through interaction [51]. Relationships should therefore be analyzed in terms of the amount of time spent interacting, the temporal sequence of those interactions, and their distribution both on common as well as individual scale, thus providing means for measuring human relations and quantitative research in the field [166, 51].

It can be argued that there is a high correlation between spatial proximity and social links [117]. In fact, social proximity has been found to be the most important feature for the detection of social interaction [144]. There is potentially a certain set of universally applicable rules for (appropriate) behaviour with respect to proximity [130, 166]. Since assuming postures, position and orientation tends to happen unconsciously, they serve as reliable cues for the attitude of people towards a social situation [333]. Additional important aspects of behaviour in regard of proximity are inclusion versus exclusion, face to face versus parallel orientation, or congruence versus incongruence [333, 336]. These can be accurately described in terms of *interaction geometry* [123, 122, 278, 67].

1.2.1 *Proxemics*

The study of human proxemics, i. e. their spatial and territorial behaviour, dates back to the early 1960s, and is founded on the seminal works of Hediger and Hall [140, 130, 131]. Hedinger found that the behaviour of animals upon contact with other animals of the same or different species depends on distance. Following his findings he established the terms flight, critical (or attack), personal, and social distance [140]. Flight and critical distance are crucial upon contact with different species. They correspond to invisible margins which, once crossed, determine whether an animal would flee or potentially attack. The latter two distances correspond to intra-species contacts and define the limits of intimate



(a)

(b)

Figure 1.: Schematics of Hall's model of personal zones (a) and Kendon's F-formation systems (b), for which the orange lines indicate the boundaries of the individual transactional segments.

and communicative distance for non-contact species, i. e. those species that do neither foster nor tolerate touch. This is augmented by Sommer who states that territory differs from personal space in that personal space moves along with the individual, whereas a territory is stationary [307]. Also, the boundaries of a territory are usually clearly marked as such, yet for personal space they are invisible. At last, while a territory is most likely defended against intruders, intrusion into personal space tends to cause withdrawal.

Hall subsequently investigated the personal space of humans [133], dividing it into intimate, personal, socio-consultive and public zones (see figure 1a). The *intimate zone* ranges from 0 to $\sim 45\mathrm{cm}$ and is typically reserved for interactions with family or close friends. The *personal zone* extends from $\sim 45\mathrm{cm}$ to $\sim 1.2\mathrm{m}$. It corresponds to the distance that people assume e. g. when talking with friends or colleagues. Dosey et al. further describe the personal zone as a buffer zone whose main purpose is the protection of the emotional well-being [78]. The *socio-consultive zone* extends from $\sim 1.2\mathrm{m}$ to $\sim 2.4\mathrm{m}$ and allows for interactions in a professional context, such as talking to a superior at work or consulting with a lawyer. The *public zone* eventually includes all interaction beyond $\sim 2.4\mathrm{m}$, for example when attending a public event or a lecture. Apart from this "semantical" meaning, Hall attributes the gain or loss of important sensory input such as olfactory or thermal perception, sight, loudness or touch to variations in distance [133]. He furthermore acknowledges that the specific extents of the four zones apply to western caucasian Americans, and may vary with additional parameters like culture or ethnological heritage (refer to section 2.4.1 for further details). He is also aware that "social organization is a factor in personal distance" [133], which shows, for instance, in that impersonal business is conducted at greater distances than when working together, independent of social distance.

Summing up, from a sociopsychological perspective personal space can be interpreted as a functional, mediating, cognitive construct which allows the human organism to operate at acceptable stress levels and aids in the control of intraspecies aggression [90]. A more detailed overview on the history of proxemics can be found in [22].

1.2.2 *F-Formations*

Kendon later criticized that most former research was concerned with individuals rather than with systematic and behavioural formations of multiple subjects [166]. According to Kendon, during interactions people are arranged according to geometric patterns, which can vary in their nature from static to highly dynamic. Static arrangements are referred to as formations, for which he assumes that although every encounter is unique per se, they all share universally applicable principles. According to his seminal work [166], a so-called *F-formation* occurs whenever two or more people form a spatio-orientational relationship. The contextual system that leads to this formation is consequently termed F-Formation System (FFS). Note that a FFS is independent of its participants as individual subjects. Instead, the system depends on their contextual relations. Therefore a FFS may remain stable even though individual members are exchanged. As an example, one person standing in a circle together with others could leave that circle, upon which the remaining members might adopt another formation while the system per se stays intact. Kendon's research is further motivated by the following insights:

- FFSs function as the identity and integrity of any social interaction.
- In spite of their common focus, FFSs allow for different ways of interaction.
- FFSs form a unit for social encounters. As such, they have a bounding or limiting effect.
- FFSs yield a spatial organization of behaviour within a social situation.

Space in every formation is partitioned according to any participating person's Transactional Segment (TA), Object Space (O-space), Personal Space (P-space) and Rear Space (R-space) (see figure 1b). In regard of the TA, recall that an individual's activity is always related to space. The space in which an individual acts is therefore called their TA, and people try to maintain that segment as long as they perform any corresponding transactions. According to Kendon, "others respect this space, not entering it or crossing it." [166]. The layout of the TA depends on location and orientation of the body. Respective changes are therefore immediately reflected in the individual's primary line of activity. Kendon relates body orientation specifically to the orientation of the lower body because it constrains the movement of the upper body, while head and arms move freely. The intersection of the individual TAs defines the O-space. The O-space is therefore always located in front of a person, and its presence is a prerequisite for when people act together. Their coordinated efforts then lead to establishing the O-space. Note that the existence of an O-space is sufficient for considering any F-formation as being fully established. P-space denotes the portion of space that is occupied by the subjects' bodies. R-space corresponds to the space that is not accessible by the individuals, and usually refers to their rear. Interestingly enough, R-space may act as a buffer zone which e. g. may be relevant when two or more groups try to establish compatible arrangements in a confined environment. If it is not possible to avoid the buffer zone, body language, such as looking down or away, is typically used to communicate respect or lack of interest [166]. In addition to the aforementioned spaces, the so-called Face Address System (FAS) accounts for the fact that people look at the persons they are speaking or listening to. In most cases, FAS and TA intersect each other, although there may be situations where people briefly address somebody outside of the TA. When it turns out that the latter may last (unexpectedly) long, the TA will shift accordingly.

1.2.3 *Properties of Spatio-Orientational Arrangements*

Spatial and orientational arrangements of multiple persons are versatile. People usually adopt circular, semi-circular, rectangular or linear formations [166]. Two persons, for example, could be standing in a face-to-face configuration, be arranged in the shape of an "L", or alongside each other while looking into same direction [167, 166]. Among other things, the selected arrangement depends on the number of persons, spatial constraints or a common activity. Additional constraints may exist due to the presence of other individuals or nearby F-formations, which will usually respect each other. Arrangements are furthermore influenced by sociofugal or sociopetal forces due to architectural factors, furniture or the placement of objects [224]. Vice versa, any concrete arrangement can also convey information about the group, e. g. whether they are acting in contest, working together or alone, or about attributes such as dominance and social hierarchy [305, 306, 61]. Circular arrangements, for example, may indicate equality among a group's members, whereas individual shifts from commonly adopted arrangements may indicate more "weight" in a person's role, e. g. when a group of students were talking to their professor. Orientation is therefore an important addition to Hall's model of personal distances. It follows that shifts or changes in the arrangements hint at underlying organizational or hierarchical changes. Other than that, as a FFS may also exist to support a certain utterance exchange, it may also shift along with a topic change, especially in situations where people are standing [166]. Once established, groups however try to maintain their arrangements. As a consequence, individuals move along with each other, i. e. they work together towards sustaining a FFS. The forward movements of one person might for instance be compensated through the backward movements of another person [28]. Goffman refers to this behaviour as a *working consensus* [114], where the system is kept in equilibrium. Note that, naturally, for every person their participation in a social situation yields their subjective affective meaning. Therefore people instinctively try to establish and maintain a common affective meaning as soon as they come together [286, 250, 198]. Triandis furthermore distinguishes between an individual's private, public and collective selves, each of which pursue specific goals [325]. In particular, it is the public and the collective selves who have the desire to act appropriately and be rewarded through corresponding backchanneling, and make efforts towards achieving the common interests of the peers [325].

At last, note that changes of an arrangement do not necessarily imply a change or the end of the respective FFS. Sometimes arrangements simply adapt to changes in the environment. Also, adjacent FFS constantly influence each other despite the efforts of their participants to uphold formations and interactions. Another noticeable shift may be due to “lurkers”, i. e. persons not (yet) actively participating in a social situation. Once outsiders approach an established system, they will typically stop at some outer position to show that they respect the boundaries of the system, yet also signal their wish to enter. When the group opens up, the arrangement will adopt the newcomer and then stabilize again once (subtle) salutations were changed.

1.2.3.1 *Where does interaction start and where does it end?*

In general it is difficult to tell exactly when social interaction starts and when it ends. Behaviour is not discrete but continuous. Behavioural cues, social signals, actions and interactions may or may not be hierarchically organized and can be regarded at different levels of abstraction. When two people meet, for example, does interaction start once they establish eye contact or when they shake hands? Although certain behavioural phases go along with respective variations in posture and distance [283], it is not possible to find a total ordering of events. Hence the question is whether any meaningful boundaries can be defined, and whether these could apply to different types of social interactions or, more generally, social situations. For the given example, Kendon initially attempted to identify the earliest time at which one could speak of greeting behaviour [166], but found that gestures, gaze etc. follow various patterns even though one might assume that a greeting scenario would be rather restricted in terms of available patterns and their interpretation. Sometimes, certain behavioural cues appear, but the same do not appear at other times. Interaction however clearly occurs when the behaviour of one person *observably* depends on that of another [166, 336]. The decision process should thus begin at the most inclusive level, i. e. when interaction is clearly observable and agreeable, and from there the process should continue outwards. Spatial and orientational arrangements can hence serve as indicators for the beginning and ending of social interaction, and changes of arrangement can relate to changes in the kind of interaction. For research of social interaction in the context of SSP it follows that fuzzy boundaries between interaction and non-interaction are acceptable, provided that fully observable interactions are clearly separable from non-interaction, and that they can be agreed upon.

1.3 RESEARCH QUESTIONS

Following the prior discussion, SSP is comprised of the analysis, modeling, and synthesis of non-verbal social interaction. In the context of SSP, this work is concerned with the question for suitable means of capturing social context on small spatio-temporal scales through the use of mobile agents, and how that context can be modeled and characterized.

This should furthermore be realized in a way such that the mobile agents will not depend on external infrastructure.

Social relationships constitute an elementary aspect of the context of a social situation, and are quantifiable as functions of social interaction. Based on the detection and interpretation of corresponding social signals and behavioural cues, this work shows how actual social situations can be inferred from social interaction. For this, a social situation is defined as co-located face-to-face social interaction with full mutual awareness of all participating persons. As such, it is denoted as a four tuple $\mathsf{S} = (\mathsf{P}, \mathsf{T}, \mathsf{X}, \mathsf{K})$ of P a set of persons, $\mathsf{T} \in \mathbb{R}$ a temporal reference, $\mathsf{X} \in \mathbb{R}^3$ a spatial reference, and K a set of tags which can be used to describe the situation's semantics. Note that T and X are actually projections from a spatio-temporal reference $\tilde{\mathsf{X}} \in \mathbb{R}^4$ to account for shifts of location over time. Also note that full mutual awareness of all participants implies the deliberate exclusion of potential overlaps with other situations.

The first research question is based on the theory of proxemics and F-formation systems. It investigates the realization and quality of a new algorithmical model for the detection of social interaction based on behavioural cues from dyadic interaction geometry corresponding to interpersonal spatio-orientational arrangements.

The second research question investigates the potential effects of personal profile parameters and additional latent variables onto the model, and whether and how they could be integrated into the process.

The third and fourth research questions concern the automatic measurements of location and orientation of mobile agents, and how such measurements can be related to the actual location and orientation of the human body.

The fifth research question investigates the fusion of physical and logical sensors from one or more agents, along with modeling and integration of subjectivity and mutual trust, in order to infer n-ary social situations from dyadic social interaction.

The last research question investigates the feasibility of a new model for dynamic social interaction based on the detection of simultaneous co-located identical activities as history-based estimates of social interaction. The model should detect universal activities of the same type and not be constrained to an *a priori* determined set of activities.

2 SOCIAL INTERACTION GEOMETRY

2.1 INTRODUCTION AND RELATED WORK

In 2009, Amoaka et al. developed a basic probabilistic model of personal space for use in computer vision supported Human Computer Interfaces (HCIs), e. g. for applications of virtual agents in public spaces [13]. In accordance with Shozo, [299] in [13], they assumed personal space to be twice as wide in front of a person than in their rear. Their model is consequently comprised of two multivariate Gaussians, centered around the person's head, and with trivial diagonal covariance matrices aligned according to the direction into which the person is looking. All parameters of the covariance matrices depend on the standard deviation of a single variable along the horizontal axis. In its basic form, the model therefore only has a single degree of freedom. This is somewhat compensated by subsequently lifting this variable to a function of three personal profile parameters, namely gender, age, and a third individual parameter, which altogether act as a linear (or potentially non-linear) filter. At the bottom line, it is interesting to see that Amoaka et al. already considered the integration of profile parameters into models of personal space. On the other hand, their model is rather artificial as it is based on manually designed and overly simple probability distributions, instead of e. g. being inferred from statistical quantities. To the best of the author's knowledge, the publications of Shozo [299] and Amoaka et al. [13] were the only preliminary works in advance of the studies leading to the first parts of this chapter, published by Groh et al. in [123].

Cristani et al. subsequently developed a robust computer vision algorithm for the detection of F-Formation Systems (FFSs) [66]. In a first step, their system determines the positions and head orientations of the subjects. In a second step, a voting mechanism leads to the identification of O-spaces, which is sufficient for the presence of established FFS [166] (see section 1.2.2). A last step verifies that no other subjects are located inside a candidate O-space. One especially interesting part of their approach is the integration of uncertainty into the model. For this, multiple samples are drawn from a multivariate distribution around the location and orientation of a subject's head. Votes are then computed for each of the samples on a discrete grid of possible centers of O-spaces. As a consequence, their model exhibits robust performance on both real-world and synthetic datasets [66, 65].

In a subsequent related work [67], Cristani et al. have shown how interpersonal distance correlates with social relationship, for which they monitored 13 subjects in casual standing conversations from a bird's eye perspective using a single fixed camera. They argue that the correlation between social relation and interpersonal distance is higher than that of social relation and orientation, also citing [111]. Based on their aforementioned algo-

rithm for statistical analysis of FFSs [66], beginnings and endings were determined for any stable FFSs, considering only those FFSs that lasted longer than five seconds. For every such formation, the pairwise distances of each adjacent pair of persons were determined. Subsequent Expectation Maximization (EM)-based clustering then revealed that all measurements were distributed among three to five modes with normal distributions. As a result of their experiments, Cristani et al. were able to relate the pairwise measurements in any of these modes to the apparent social distances between the respective people (professors, PhD students, undergraduates). They furthermore showed that the means of the modes would adapt to additional constraints when imposed on the room in which the subjects could freely move, but that they would still adhere to the same number and prior distribution of clusters [67].
Around the same time, Hung and Kröse proposed the estimation of FFSs through dominant sets, a "form of maximal clique that can be applied to edge weighted graphs so that the affinity between all nodes *within* [the subgraph] is higher than between the internal nodes and those that are external to it" [149]. In their work, *affinity* is computed based on relative distance, orientation, a custom feature called Socially Motivated Estimate of Focus Orientation (SMEFO), or combinations of the former. One may note that, whereas distance-based affinity is modeled by the exponential of Euclidean distance weighted by the function's variance, orientation is only trivially integrated into the model by cropping the distance-based affinity to zero if at least one person A in a pair (A,B) is oriented such the other person B is not located in A's frontal hemisphere. The proposed SMEFO feature, however, follows the presumption that people who attempt to interact stand more closely together and orient themselves accordingly. SMEFO therefore depends on the angle of the vector from a person's position to their estimated center of focus, the latter of which denoting the weighted sum of distance-based affinities towards all other persons.
Out of a comparatively large dataset, for which 50 persons were recorded from a bird's eye camera over the course of three hours, 82 images comprised of ~ 1700 persons were annotated with location and orientation of each subject [149]. In addition to that, a group of human experts, notably from different cultural backgrounds, labeled the apparent FFSs in overlapping sets of images, yielding an agreement of more than 94%. It should be pointed out that, although human experts will undoubtedly take into account more than just proxemic behavioural cues when labeling still images of social situations, this very high agreement contributes to the argument that location and orientation are indeed significant priors for the existence of FFSs, and therefore social situations. It furthermore sustains the assumption of a basic subset of rules for proxemic behaviour that may be universally applicable among humans. Readers should note the emphasis on *basic*, as it is already known from social sciences that proxemics are influenced by additional parameters [130, 133, 166]. All in all, Hung and Kröse report good results for the positive detection of FFSs when only distance-based affinities where used. Augmenting those distance-based affinities with SMEFO would sometimes yield small improvements, but it appears that this would not be the general case. High precision and recall were yet achieved for the combination of distance- and orientation-based affinities. Setti et al. proposed a revised formulation of [67], relaxing the constraints imposed on the prior model of O-space based

on a single multivariate Gaussian through the use of an entropy based voting mechanism with respect to varying group cardinalities [291]. More precisely, $K - 1$ voting modules are employed for cardinalities $k \in \{2, \dots, K\}$ on an image of K persons, each with respect to circular arrangement of the k subjects with an assumed distance of 95 cm between adjacent persons, accounting for placement within the personal zone. Each module produces weighted entropy measures based on the times and weights of the subjects' votes for potential centers of O-spaces. The accumulated voting spaces for each cardinalty k are then pruned of all those candidate O-spaces with differing k. The results are eventually merged in a multi-scale accumulator from which, for every discrete location, the FFS with the highest entropy is selected. According to Setti et al. [291], their revised approach outperformed all prior attempts of statistical analysis of FFS in still images, including the aforementioned [67] and [149], in most cases demonstrating considerably higher precision and recall.

Altogether, related work shows that designed, constrained or simplistic models suffer from a lack of expressiveness for human proxemic behaviour. This does, however, not necessarily imply a demand for sophisticated models, a fact clearly shown by the considerable results that were achieved based on those models that involve clever voting mechanisms, incorporate elementary findings from the sociological theory on FFSs, and/or employ means of uncertainty. The related work furthermore shows that the integration of orientation into the decision process improves the quality of the results and allows for decisions in situations where interpersonal distance alone would not be sufficient. It is also clear that interactants arrange themselves in various formations, depending on sociopetal or sociofugal forces (see section 1.2.3), environmental constraints, and most importantly social factors such as relationships or the affective meaning of the situation. Computer vision based algorithms could, without doubt, automatically recognize and handle several constraints, such as e. g. obstacles in the environment. To a limited extent, these or similar types of constraints may even be integrated into the model itself, yet only for specific applications at known locations. Aside from environmental constraints of a more *static* character, the presence of other nearby FFSs, together with their specific arrangements, can be regarded as *dynamic* constraints. So far, however, the aforementioned models take into account neither static nor dynamic constraints. Instead, the models have a local focus on each distinct FFS, except for the model by Setti et al. [291], whose multi-scale voting for every individual subject in the scene implicitly accounts for multiple FFSs on a more global scale. Still, even the latter model is potentially restricted due to the fact that it is based on heuristics such as the presumption of circular arrangements and/or typical distances of 95cm between adjacent persons. Instead of the explicit integration of dynamic constraints, and irrespective of possible extrema, models would more likely benefit from implicitly learned knowledge, such as is the case for quantitative models.

The present work thus proposes the algorithmic detection of social interaction based on quantitative models learned from measurements of interaction geometry in dyads. For this, interaction geometry will be modeled as a triple $(\delta\varphi, \delta\theta, \delta d)$ of interpersonal distance (δd), relative orientation $(\delta\theta)$, and relative location $(\delta\varphi)$. So far, relative location has not been considered by other models, although clearly, interaction geometry is only fully determined once $\delta\varphi$ is taken into account, as distance δd merely accounts for infinitely many positions on a circle and a description of orientation $\delta\theta$ is per se independent of any other variable.

The layout of this chapter is as follows: Section 2.2 describes the aquisition, post-processing and annotation of a sufficiently large dataset for social interaction, which is subsequently analysed and discussed in terms of the aforementioned variables of interaction geometry. Section 2.3 provides a detailed derivation and evaluation of the resulting algorithmic model for the detection of social interaction. Following the discussion that social interaction is potentially influenced by further variables, such as, for instance, personal profile parameters or the cardinality of groups, section 2.4 discusses influential factors in general, and features a second dataset as well as a corresponding model in order to evaluate the actual correlations of gender and age with respective measurements of interaction geometry.

2.2 EXPERIMENTAL DATASET OF SOCIAL INTERACTION GEOMETRY

The dataset at hand is based on the recordings of an experiment which was conducted at the computer science department of the Technische Universität München on December 21st, 2009. During this experiment, position and orientation of the participating persons were continously monitored by an infrared tracking system over the course of 30 minutes. The subjects were furthermore recorded by stationary as well as mobile video cameras. Audio was captured for each of the interactants through small wearable recording devices. In order to gather a preferably rich and diverse set of data, in particular comprised of multiple naturally changing as well as lasting social situations and varying group cardinalities, the participants were instructed to determine each other's favorite food, TV show and music at childhood. As an additional incentive, participants could win a price valued at about 30,- EUR, provided they would give the quickest correct answers when asked about the favorites of other, randomly selected, members of the group. Of the 9 participants, 7 were male and 2 were female. Age ranged from 22 to 31 with a median of 23, mean of 24.77 and standard deviation of 3.19 years. Body height ranged from 165 cm to 186 cm with a median of 173 cm and mean of 175.1 cm. According to [2], the present time average height of German males and females is 178 cm and 165 cm. The respective standard deviations for male and female experimental subjects were 5 cm and 8 cm. All but one person were of native German origin, the exception being one student with Asian heritage. The following sections give a detailed description of the recording, post-processing, annotation and final analysis of the data.

2.2.1 *Recording*

2.2.1.1 *Video Cameras*

Throughout the experiment, the participants were filmed by 6 stationary high-resolution cameras, plus one backup mobile camera. The stationary cameras were all mounted on the ceiling such that they would record the scene from various angles. Likewise, the mobile camera was only used to record the scene from the "outside", so that none of the cameras would interfere with the subjects' behaviour inside their moving area. During postprocessing, the mobile camera could provide more detailed information about certain formations when obfuscated from the stationary cameras through motion, position or mutual shadowing of the participants, e. g. due to body height. All cameras provided digital video streams at 25 frames per second, stored in a custom container format including precise time stamps. Eventually, the video streams were precisely aligned with the time scale of the infrared tracking system, henceforth aiding in the clarification of the exact set of persons during the subsequent annotation of the monitored social situations.

2.2.1.2 *Infrared Cameras*

Position and orientation of the participants were tracked using a system of 8 stationary infrared cameras [8], 4 of which were mounted on the ceiling and 4 on the floor. The system is capable of tracking up to 20 markers in real-time, for which each marker features a number of spheres with a specular surface that reflects the infrared beams sent from the cameras at a rate of 60 Hz. The number and configuration of the spheres are unique for each marker, so that each target can be detected without ambiguity. Each subject wore a single marker on either their left or right shoulder, located at a distance of 18 centimeters from the center of the torso when projected onto the x/y plane. The markers were fixated to make sure they would not skid or move throughout the process. For each marker, position and orientation in the camera coordinate system were continuously computed via metric reconstruction of sets of corresponding points [136] as seen by two or more of the cameras for every frame, for which the camera coordinate system had initially been calibrated such that its three axes were known precisely (see figure 2). These data were available at all times via a real-time TCP/IP stream [8].

2.2.1.3 *Audio Recordings*

Audio was recorded to provide additional means of control for post-processing, as well as for improving the detection of social situations through fusion of multiple sources of information (see section 4.6). For this, only the presence or absence of conversational audio would be taken into account, but not the semantics of what was spoken. Mono-channel audio was recorded through small wearable recorders, each of which taped to the breast of the respective interactant. Audio recordings were performed at a sampling frequency

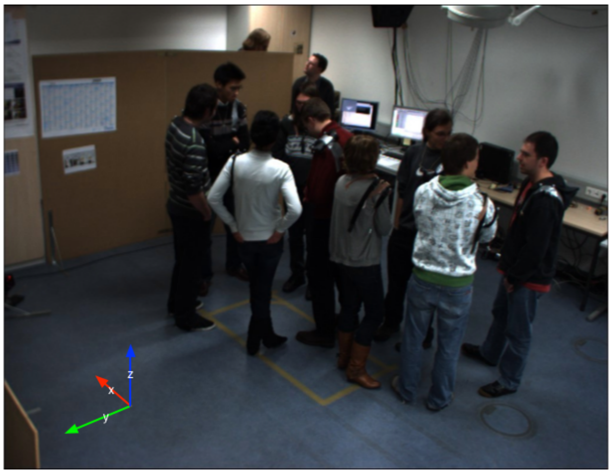


Figure 2.: “Action shot” of the recording plus a visualization of the camera coordinate system setup.

of 11 KHz. In advance of the actual experiment, a sequence of three sinusoidal signals were emitted through a common loudspeaker at 440 Hz, 880 Hz and 1760 Hz to allow for temporal synchronization of the devices.

2.2.2 *Post-Processing*

The infrared tracking system recorded a total of 121,447 frames at 60 Hz over the course of 33 minutes and 44 seconds. For each frame, the identifiers, positions and orientations of the visible markers were stored. Positions are stored as 3-tuples, specifying the x-, y- and z-coordinates of the respective marker in relation to the camera coordinate system. Orientations are stored both as 3-tuples of Euler angles as well as Direction Cosine Matrices (DCMs), for which the order of rotations is defined as subsequent rotations around the z-, y-, and x-axes [8]. Note that a representation in terms of Euler angles is prone to suffer from singularities, known as *Gimbal Lock*, which occur whenever a prior rotation aligns the two successive rotational axes, hence leading to the loss of one degree of freedom. As a consequence, only the DCMs were used throughout the following process.

Post-processing needs to be performed to clean the data of redundant and/or misleading information originating from actually unassigned, yet erroneously detected, markers, e. g. due to accidental reflections caused by clothing. From the present dataset, 2 out of the 11 detected markers proved to be false positives that showed up from time to time, which is consistent with the actual number of 9 participants. The respective markers could easily be identified and were consequently removed from the dataset. Furthermore, positional and orientational data may be temporarily missing for changing sets of markers, e. g. caused by participants shadowing each other or accidentally walking out of the area visible to the infrared cameras. Given the rather static character of the monitored FFSs and hence the temporal stability of the subjects’ spatio-orientational configurations, missing data

Marker ID	3	5	6	8	9	10	11	12	13
Frames	11100	28967	629	1579	2332	2881	5511	969	267
	9%	23%	0%	1%	1%	2%	4%	0%	0%

Table 1.: Marker availability (missing frames).

can be easily compensated through interpolation. This was further verified by detailed analysis of the respective portions of the video footage. Out of the 9 actual markers, most markers had none or at most a few missing frames. However, 2 of the 9 markers showed noticeable losses of 9% respective 23% of the total number of frames (refer to table 1). Here, analysis of the video footage revealed that marker 3 was frequently shadowed by taller persons, whereas the person wearing marker 5 stood at the margin of the observable area for a few minutes, plus the subject's long hair occasionally covered the shoulders and thus the marker. Also note that missing frames were mostly not sequentially related, such that e. g. the interpolation of the missing data for marker 3 would not correspond to a *continuous period* but to a *total* of 11100 frames $\cdot\, 3600$ frames/s $\approx$ 3 min, consisting of several sequences of variable length throughout the whole duration of the recording.

2.2.2.1 *Position*

Interpolation between the last known position at t_0 and the next known position at t_1 is straight-forward via

$$\mathbf{s}_i^t = \mathbf{s}_{t_0} \cdot (1-u) + \mathbf{s}_{t_1} \cdot u \,, \tag{1}$$

where $t_0 \leqslant t \leqslant t_1$ and $0 \leqslant u = \frac{t-t_0}{t_1-t_0} \leqslant 1$.

2.2.2.2 *Orientation*

Linear interpolation is not applicable for DCMs. DCMs form the so-called special orthogonal group $\mathsf{SO}(3)$, from which it follows that the determinant of each DCM is precisely $+1$, the columns respective axes of each rotation matrix are orthonormal and hence $\forall \mathbf{R} \in \mathsf{SO}(3)$: $\mathbf{R}\mathbf{R}^\mathsf{T} = \mathbf{I}$, i. e. the inverse of each element is simply given by its transpose. These properties are likely violated by linear interpolation. Compensating for missing orientational data is however easily achieved through quaternion algebra, which provides additional means of representing rotations and corresponding operators in $\mathbb{R}^3$ [176, 297, 68]. For this, the recorded DCMs need to be mapped to and from quaternions as follows.

MAPPINGS Aside from the notion of quaternions as hyper-complex numbers of the form $\mathbf{q} = q_0 + iq_1 + jq_2 + kq_3$ along with the rule $i^2 = j^2 = k^2 = ijk = -1$, quaternions can also be interpreted as the sum of a scalar and a vectorial part

$$\mathbf{q} = q_0 + (q_1, q_2, q_3)^\mathsf{T} \,. \tag{2}$$

It can be shown [176] that *unit quaternions*, i. e. quaternions subject to $\sum_i q_i^2 = 1$, are well-suited for the representation of rotations, in which case the scalar (real) part relates to the cosine of the half angle and the vectorial (imaginary) part to the axis of rotation. Those quaternions whose scalar part equals zero are called *pure quaternions*.
The rotation operator for a counter-clockwise rotation around the angle and axis as represented by a given unit quaternion $\mathbf{q}$ is then defined as

$$\mathcal{L}_{\mathbf{q}}(\mathbf{v}) = \mathbf{q}\mathbf{v}\mathbf{q}^* \, , \tag{3}$$

where $\mathbf{v}$ represents a three-dimensional vector $(x, y, z)^T$ in form of a pure quaternion, i. e. $\mathbf{v} = 0 + ix + jy + kz$, and $\mathbf{q}^*$ denotes the complex conjugate of $\mathbf{q}$. Now, since the product $\mathbf{p} * \mathbf{q}$ of two quaternions $\mathbf{p}$ and $\mathbf{q}$ can itself be defined in terms of the scalar product and cross product as

$$\mathbf{p} * \mathbf{q} = \left(p_0 q_0 - \sum_i q_i p_i \right) + (p_0 \mathbf{q} + q_0 \mathbf{p} + \mathbf{p} \times \mathbf{q}) \, , \tag{4}$$

equation (3) can just as well be written in matrix notation as

$$\mathcal{L}_{\mathbf{q}}(\mathbf{v}) = \begin{pmatrix} 2q_0^2 - 1 + 2q_1^2 & 2q_1 q_2 - 2q_0 q_3 & 2q_1 q_3 + 2q_0 q_2 \\ 2q_1 q_2 + 2q_0 q_3 & 2q_0^2 - 1 + 2q_2^2 & 2q_2 q_3 - 2q_0 q_1 \\ 2q_1 q_3 - 2q_0 q_2 & 2q_2 q_3 + 2q_0 q_1 & 2q_0^2 - 1 + 2q_3^2 \end{pmatrix} \cdot \begin{pmatrix} x \\ y \\ z \end{pmatrix} \, , \tag{5}$$

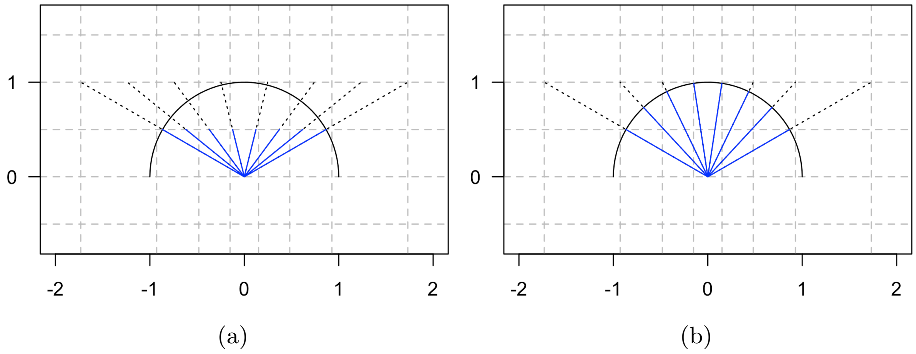
yielding a mapping from a rotation $\mathbf{q}$ in the quaternion domain to a corresponding DCMs. The reverse mapping from a given DCM $\mathbf{R} \in \mathbb{R}^{3\times3}$ to the quaternion domain can as well be derived from equation (5), by first solving the trace of $\mathbf{R}$ for the scalar q_0, and subsequently using q_0 in order to solve for the remaining vectorial parameters. More precisely, solving for q_0 yields

$$\operatorname{tr}(\mathbf{R}) = 4q_0^2 - 1 \Rightarrow q_0 = \frac{1}{2}\sqrt{\operatorname{tr}(\mathbf{R}) + 1} \tag{6}$$

and

$$\begin{aligned} q_1 &= R_{32} - R_{23}/(4q_0) \\ q_2 &= R_{13} - R_{31}/(4q_0) \\ q_3 &= R_{21} - R_{12}/(4q_0) \, . \end{aligned} \tag{7}$$

It should be noted that both equations (6) and (7) may cause problems whenever the angle of rotation is very close or equal to $180°$. Recall that q_0 denotes the cosine of the half angle of rotation and hence $\lim_{\alpha \to 180} \cos\frac{\alpha}{2} = 0$. This is clearly a problem with respect to the denominator in any one of the equations in (7). In case of equation (6), however, the problem is due to numerical cancellation and hence potentially negative radicands. Geometrically speaking, both problems can be interpreted as the fact that the axis of rotation could point into either one of two strictly opposite directions [176], depending



(a)

(b)

Figure 3.: Basic linear interpolation as opposed to spherical linear interpolation. The former lacks constant rotational speed due to varying lengths of the arcs in every segment.

on whether the rotation is clockwise or counter-clockwise. This is a well-known issue, and various methods for the extraction of a quaternion from a DCM do exist [172, 127, 93]. Apart from e. g. case-by-case analysis of the signs and magnitude of the vectorial elements, both angle and axis of rotation can be easily determined through solving the eigenvalue problem for $\mathbf{R}$. Since all rotation matrices have the eigenvalues $+1, e^{+i\theta}, e^{-i\theta}$ and since the trace of a matrix is equal to the sum of its eigenvalues, the angle of rotation can be determined as follows:

$$\begin{aligned} \operatorname{tr}(\mathbf{R}) &= 1 + e^{+i\theta} + e^{-i\theta} \\ &= 1 + \cos\theta + i\sin\theta + \cos\theta - i\sin\theta \\ &= 1 + 2\cos\theta \\ \Rightarrow \theta &= \arccos\frac{\operatorname{tr}(\mathbf{R}) - 1}{2} \end{aligned} \tag{8}$$

The axis of rotation is then simply the eigenvector corresponding to the eigenvalue $+1$, following from the fact that points along this eigenvector will be neither changed nor scaled by any rotation, according to the eigenvalue equation $\mathbf{R}\mathbf{x} = \lambda\mathbf{x}$.

SPHERICAL LINEAR INTERPOLATION Mapping to the quaternion domain has not completely solved the problem yet, since linear interpolation of two unit quaternions lacks constant rotational speed as the rotational segments vary in length (refer to figure 3). [297] therefore introduced the SLERP algorithm which solves this issue through *spherical linear interpolation*, expressible as the *quaternion product*

$$\mathbf{R}_i^t = \mathbf{q}_{t_0} * \left(\mathbf{q}_{t_0}^{-1} * \mathbf{q}_{t_1}\right)^u \tag{9}$$

or, both easier and more efficient in terms of quaternion operators, as

$$\mathbf{R}_i^t = \frac{\sin((1-u)\theta)}{\sin\theta}\mathbf{q}_{t_0} + \frac{\sin(u\theta)}{\sin\theta}\mathbf{q}_{t_1}\ , \tag{10}$$

where θ denotes the angle between the quaternions $\mathbf{q}_{t_0}$ and $\mathbf{q}_{t_1}$.

2.2.2.3 *Mapping marker position and orientation to their respective body counterparts*

At this point, the process of cleaning and compensating for missing data yields positions and orientations for each *marker* at every point in time throughout the whole recording. Recall that positions are always given as a three-tuple of x-, y- and z-coordinates in millimeters, and orientations are given as DCMs. All coordinate systems are right-handed. The infrared system's manufacturer [8] defines the transformation of a point v^M from the local marker coordinate system (M) into the global camera coordinate system (C) as

$$v^C = R_{i,t}^{MC} \cdot v^M + s_{i,t}\ , \tag{11}$$

where, for a given marker i and time $t \geqslant 0$, $s_{i,t} \in \mathbb{R}^3$ yields the marker's position, and the columns of $R_{i,t}^{MC} \in SO(3)$ correspond to the images of the axes of the marker coordinate system in camera coordinates, equivalent to a rotation which aligns the axes of the camera with the axes of the marker. An alternative view of the rotation described by $R_{i,t}^{MC}$ as a sequence of rotations with respect to the marker's *initial orientation* at $t = 0$ per

$$R_{i,t}^{MC} = \hat{R}_{i,t}^{MC} R_{i,0}^{MC} \tag{12}$$

allows for the definition of the marker's *relative rotation*

$$\hat{R}_{i,t}^{MC} = R_{i,t}^{MC} \left(R_{i,0}^{MC}\right)^{-1} \overset{\textit{orthonormal}}{=} R_{i,t}^{MC} \left(R_{i,0}^{MC}\right)^{T}\ . \tag{13}$$

Note that, for the present dataset, the initial orientation $R_{i,0}^{MC}$ was determined for each marker by averaging and consequently orthonormalizing the respective rotation matrices over the first 500 recorded frames. During this time, all participants were instructed to stand still with their upper bodies parallel to the x-axis of the camera coordinate system and facing the room's rear wall, thus looking into the direction of the negative y-axis (refer to figure 2).

In addition to the marker (M) and camera (C) coordinate systems, let the body coordinate system (B) describe the orientation of the shoulder line and the direction which the front of the body is facing (see figure 4). It differs from the marker coordinate system by two subtle but very important differences. First of all, all tracking data have been recorded for the *markers*, which is why all translations and rotations of the *body* must in fact be perceived as occuring around the marker. The location of the marker therefore represents the origin of the body coordinate system. Secondly, the axes of the body coordinate system were initially not in strict alignment with the axes of the camera coordinate system. The application of a corresponding *initial orientation correction* is therefore mandatory. As a consequence, points in body coordinates have to be transformed prior to any rotations and translations originating from the recorded data. The corresponding transformation is easily found based both on the assumption of a rigid body and the knowledge of the marker being firmly attached precisely along the shoulder line at a distance of 18 centimeters from

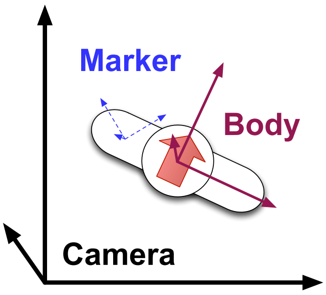


Figure 4.: The camera-, marker- and body coordinate systems used for tracking a person's position and orientation through an infrared marker.

the center of the body (see section 2.2.1.2). It follows that points in body coordinates need only be translated about that particular distance along the local x-axis, thus implicitly aligning the marker and the origin of the body coordinate system, followed by a subsequent rotation which aligns the axes of the body and camera coordinate systems.

Let $\hat{R}^{BC}_{i,t} = \hat{R}^{MC}_{i,t}$, since the orientation of body i at time t can just as well be understood as a sequence of rotations as in equation (13). The projection of body onto camera coordinates is consequently defined as the function

$$f : (i, t, v^B) \mapsto R^{BC}_{i,t} \left(R^{BC}_{i,0}\right)^T R^{IOC} \left(v^B + o_i\right) + s_{i,t} \,, \tag{14}$$

where R^{IOC} denotes the initial orientation correction and $o_i = (\pm 180, 0, 0)^T$ an offset depending on whether the marker was worn on the left or right shoulder, respectively. As stated above, all participants were initially oriented such that their shoulder lines were aligned with the x-axis of the camera coordinate system and they were facing the rear wall of the room. R^{IOC} therefore equals a constant rotation about $180°$ around the z-axis and equally applies to all bodies (hence no index i).

2.2.3 *Annotation*

The mapping from section 2.2.2.3 allows for precisely tracking and visualizing each individual's absolute position and orientation throughout the experiment. Recall that the final goal however will be the discrimination of established and non-established social situations for every participant at each point in time, based solely on interaction geometry. It is hence mandatory to annotate the dataset with the ground-truth of whether two or more subjects were interacting, and if so, who and when that was.

For this purpose, a system was developed [262] in order to allow human experts to perform the actual annotation based on an orthographic projection of the gathered data. The main component of the system is an application that sketches the whereabouts of the participants on a per-frame basis and allows for associating sets of participants with social situations. An arbitrary number of social situations can be created. Each participant cannot be assigned to more than one situation at the same time. Aside from navigating

	Group size						
	2	3	4	5	6	7	9
#	9	8	5	5	2	1	4
$\sum$ duration	1245.0s	781.2s	838.5s	536.2s	144.3s	100.5s	345.2s
Min	47.8s	22.8s	12.3s	22.2s	37.7s	100.5s	19.7s
Max	370.0s	340.7s	609.7s	245.2s	106.3s	100.5s	154.8s
Mean	138.7s	97.5s	167.5s	107.1s	72.0s	100.5s	86.1s
Median	98.5s	66.2s	80.7s	84.2s	72.0s	100.5s	85.0s
StdDev	100.6s	104.8s	250.8s	83.3s	48.6s	–	67.1s

Table 2.: Overview of the annotation results.

through and annotating mere still images, human labelers could also view the visualization as a continuous stream, which would provide additional temporal information, e. g. where otherwise it would have been difficult to determine whether a social situation had been fully established or not. In addition to that, human experts could of course also rely on the time-stamped video footage as a general fallback mechanism, especially so for the purpose of cross-checking their annotations, since from the video footage they could furthermore see many more social signals than just interaction geometry as in the projection.

The complete dataset was annotated during the proceedings of [262] and the results were thoroughly double-checked by the author of this work. All in all, 34 independent and mostly parallel social situations were identified over the course of 31:51 minutes, the first starting at frame 679 and the last ending at frame 12144. The late start is due to the obligatory calibration of the participants' markers at the start of the recording process, further explained in section 2.2.4. Table 2 shows the frequency with which social situations occurred among exactly N participants, along with detailed statistics. Moreover, figure 5 gives an overview of when and for how long social situations took place, and how many individuals participated in these situations. From these it follows that groups of two or three individuals formed more often than others, while groups of eight did not occur at all. One may note that social situations of four persons usually lasted longer than the more frequent ones with groups of two or three. Groups of six or seven persons rarely ever formed, and only during the first half of the experiment. During the second half, groups of two, three, four or five persons were the most dominant cardinalities.

2.2.3.1 *Discussion*

An important question for annotation is what is to be classified as a social situation and what is not, but also precisely where those situations start and where they end. As discussed in chapter 1, social situations can be recursively nested almost arbitrarily deep. What is understood as a social situation is often a matter of the application-specific con-

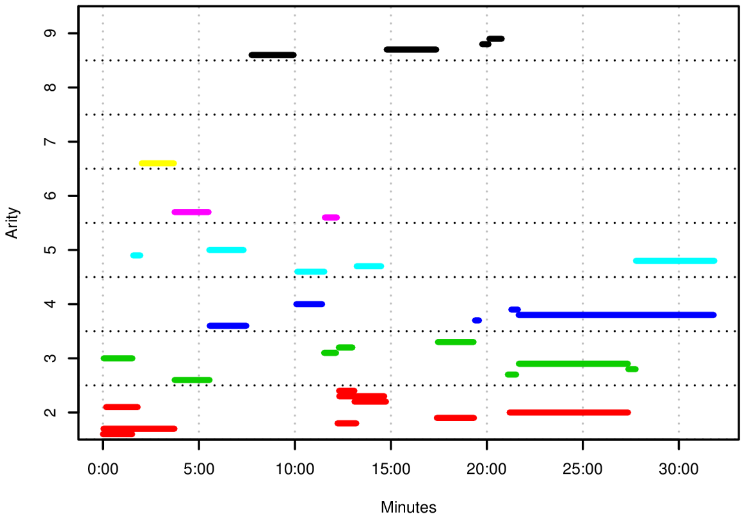


Figure 5.: Overview of when and for how long social situations took place, grouped by arity. Distinct situations with equal arity are stacked on top of each other.

text in which they are to be investigated, but certainly also dependent on a personal point of view. For example, two persons engaging in a mutual social situation can as well be regarded as a mere subset of a much greater social situation with additional persons in the very same room. According to Goffman [114], co-present persons will basically always exchange information and/or communicate, and hence they will interact, irrespective of whether they are actively or subconsciously engaged in the interaction. The issue of what should be identified as a social situation is actually resolved by the definition of social situations as given in chapter 1, according to which, in the context of this work, social situations are only considered as such in case of face-to-face interaction and full mutual and conscious awareness of all persons. Excluding potential overlaps, this yields a clear understanding of the corresponding "threshold" in a nested hierarchy of social situations. The requirement of full mutual and conscious awareness necessarily leads to the same set of persons in a social situation as seen from every single interactant.

Likewise, different approaches for determining the precise beginnings and endings of interaction were discussed in section 1.2. According to [282, 283] in [166], the "spatio-temporal frame" serves as a reliable source for identifying interaction. These frames vary upon changes in behavioural phases. Hence social interaction occurs whenever there are *observable* interdependencies between the behaviour of corresponding individuals. Erickson [89] corroborates this view by stating that social occasions, the times at which they begin, how long they last, and potentially even their whole context, are uniquely determined through (observable) parameters like speech, proxemics, orientation and posture. Moreover, according to Kendon [166], the existence of a common O-space is sufficient for the presence of a FFS, and therefore the presence of a social situation. It follows that "brackets" around social interaction should be determined from the inside out, since the transition between non-interaction and interaction is fluent and (highly) context-sensitive. During annotation, finding “brackets” for phases of clearly established social interaction is a rather straight-forward task. On the other hand, the purpose is to develop a preferably universally applicable model. In that sense, such a model can only attempt to learn the “true” ratio between orthogonal decisions for marginal cases during transitory phases, meaning that the model will only profit from a certain degree of fuzziness. For marginal cases, one could otherwise only come to the right conclusion if the social context were (fully) known. Recall, however, that behavioural cues supposedly have no intrinsic meaning at all [166, 283]. It can be presumed, though, that the mere distribution of samples of interaction geometry may still yield implicit, more complex, information about relationships, mood, culture, gender, etc. It is, for example, the relative frequency and/or the actual distribution of the samples that implicitly encodes information about time spent in, intensity of, and limited dynamics of social interaction. Any mathematical model for this must hence be built upon this contextual information. In turn, this means that potential issues will vanish along with an increase in the number of monitored social situations respective corresponding samples for both present and non-present social interaction.

At last, one may furthermore ask whether, or to what extent, annotations made by different human experts are likely to yield the exact same results. For this, recall that the annotation is not solely performed based on the orthographic projection of the recorded data, but just as well on the video footage of the experiment. Humans naturally make use of a vast number of physical and (socio-)logical sensors like their ears, eyes, smell, touch, interpretation of facial expressions, postures, head tilt, etc., for example in particular so when deciding whether a certain scene constitutes a social situation, and who precisely is part of it. In addition to the orthographic projection, the video recordings convey a lot more such sensoric information to the expert during annotation, which apparently forms a common ground for decision making. This notion is sustained by the findings of Hung and Kröse [149] which were discussed at the beginning of this chapter. In their experiments, the consensus between human annotators on a very large dataset was found to exceed 94%, even though these annotators had notably different backgrounds. Nonetheless, this matter could be further investigated by conducting experiments where subjects are provided with various spatio-temporal configurations through either video, orthographic projection, or both, and subsequently comparing the results of their annotations. It would also be interesting to see how these experts perform with respect to the aforementioned additional sensors when given orthographic projections along with either the corresponding continuous video streams as opposed to only still pictures from that video. For example, how would a scene where several people which are actually in a social situation, and where one person briefly turns to look at something outside of that situation, be classified.

2.2.4 *Variables of Interaction Geometry*

So far, the dataset provides only the position and orientation of the participants, together with the ground-truth of who interacted when and with whom. Interaction geometry, though, is based on the *relative* mutual positions and orientations of the subjects. The

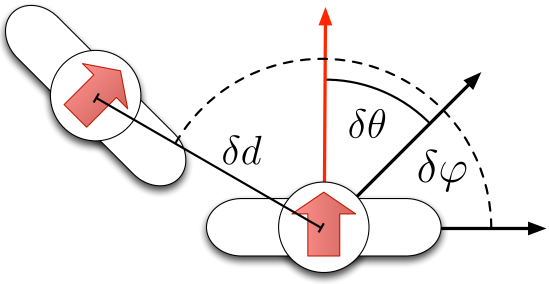


Figure 6.: Illustration of the three variables used for modeling of interaction geometry.

reason for this is two-fold: For one, interaction geometry, as a layer of abstraction, yields very good visualization and interpretability when building, understanding, and possibly adapting corresponding mathematical models. Moreover, acquisition of *absolute* measures using only mobile sensors is an extremely difficult task which, so far, has not been satisfactorily solved in research [188, 349].

Interaction geometry can be modeled for any pair of persons i and j in terms of three variables as seen by either one of i and j, namely $\delta\theta_{ij}$, δd_{ij} and $\delta\varphi_{ij}$, all expressed with respect to a right-handed coordinate system where the persons are standing on the x/y-plane and the z-axis points upwards, and

- $\delta\theta_{ij} \in [-\pi, \pi)$ describes the relative orientation of the shoulder lines, i. e. the angle about which person i must rotate around the yaw-axis such that their upper bodies are aligned in parallel and both face the same direction,
- δd_{ij} describes the relative distance between the centers of the bodies, assuming that, when projected onto the x/y plane, the center of the torso lies exactly half-way between the shoulders,
- $\delta\varphi_{ij} \in [0, 2\pi)$ describes the position of person j in relation to person i. This angle is measured between the positive x-axis of the local two-dimensional coordinate system of person i and a vector from the origin to the center of the body of person j, where the origin is located at the center of the body of person i and the x-axis is parallel to the upper body, pointing at the right shoulder.

Figure 6 provides an illustration of $\delta\theta$, δd and $\delta\varphi$. Note that whereas $\delta\theta$ and δd are symmetrical, $\delta\varphi$ is not as it depends on the orientation of the upper body of the person from whose perspective the relation is described, from which it follows that the three-tuple $(\delta\theta, \delta d, \delta\varphi)_{ij}$ is also not symmetrical. The model of interaction geometry must therefore be based on *both* observations from i to j and j to i.

2.2.4.1 *From position and orientation to variables of interaction geometry*

Computing the newly introduced variables of interaction geometry from the present data is straight-forward. As defined in section 2.2.4, $\delta\theta$ describes the relative orientation of the upper bodies with respect to the yaw-axis. This is actually equivalent to the relative rotation Q which would align the shoulder lines of persons i and j in parallel, and hence:

$$\begin{aligned} R^{MC}_{i,t}\left(R^{MC}_{i,0}\right)^{T} &= Q\,R^{MC}_{j,t}\left(R^{MC}_{j,0}\right)^{T} \\ \Leftrightarrow\quad Q &= R^{MC}_{i,t}\left(R^{MC}_{i,0}\right)^{T}\left[R^{MC}_{j,t}\left(R^{MC}_{j,0}\right)^{T}\right]^{-1} \\ \Leftrightarrow\quad Q &= R^{MC}_{i,t}\left(R^{MC}_{i,0}\right)^{T} R^{MC}_{j,0}\left(R^{MC}_{j,t}\right)^{T} \end{aligned} \tag{15}$$

The angle of rotation about the yaw-axis can be directly computed from the rotation matrix Q, which would however require knowledge of the exact sequence of rotations that finally led to the DCMs in the recorded data. In spite of the fact that [8] defines the order of rotations as x (pitch), y (roll) and z (yaw), a more general solution, which does not depend on any prior knowledge, is given by first transforming an arbitrary point v on the x-axis, say $v = (1, 0, 0)^T$, by Q, and then finalizing $\delta\theta$ as the angle between the x-axis and a vector from the origin to the transformed point, i. e.

$$\delta\theta = \text{arctan2}\left(v'_2, v'_1\right) \quad \text{where} \quad v' = Qv\,. \tag{16}$$

For the computation of δd and $\delta\varphi$, the basic idea is modeling the body in terms of a set of points describing the center of the body, the left shoulder, the right shoulder and the nose, and transforming these points as required. Note that left or right “shoulder” really refers to a point within the distance between the body’s center and the actual shoulder. More precisely, either one refers to the location at which the marker is worn, be it on the left- or right-hand side. This is sufficient because, for the validity of the dependent variables of the current dataset, only the precise distance between the center of the torso and the marker is significant, which was a controlled parameter during the experiment (18 cm). The set of points is therefore defined as

$$\mathcal{S} = \left\{(0, 0, 0)^T, (-180, 0, 0)^T, (+180, 0, 0)^T, (0, 60, 0)^T\right\}\,. \tag{17}$$

The choice of the value for the last point (nose) from the center is rather arbitrary, as it is merely used to represent the direction into which the upper body is facing. The value has been chosen mainly for visualization as well as to avoid numerical issues. Note that this vector could just as well be determined by the cross-product $y = z \times x$ of the idealized z-axis $(0, 0, 1)^T$ and actual x-axis of the corresponding body, given by the rotation matrix $R^{BC}_{k,t}$ for any marker k.

Now, the mapping f from equation (14) is used to determine δd as the Euclidean distance between the centers of the bodies i, j at time t:

$$\delta d = \sqrt{(c_{j,t} - c_{i,t})^T (c_{j,t} - c_{i,t})} \quad \text{where} \quad c_{k,t} = f(k, t, (0, 0, 0)^T) \tag{18}$$

Next, the angle $\delta\varphi$ is measured between the shoulder line of person i and a vector $d_{i,j,t} = c_{j,t} - c_{i,t}$ from the body center of person i to its counterpart of person j. As the angle is defined with respect to the local coordinate system of person i, the vector d must be transformed accordingly. For this, let $B_{j,t} = R^{MC}_{j,t} \left(R^{MC}_{j,0}\right)^{T} R^{IOC}$ an orthonormal matrix, x be a vector in camera- and y be a vector in body coordinates. Since d represents a direction, there is no need for translation, so that the transformation from camera to body coordinates follows from

$$\begin{aligned} Ix &= By \\ B^{-1}Ix &= y \\ \left(R^{IOC}\right) R^{MC}_{j,0} \left(R^{MC}_{j,t}\right)^{T} x &= y \;. \end{aligned} \tag{19}$$

$\delta\varphi$ is then determined as

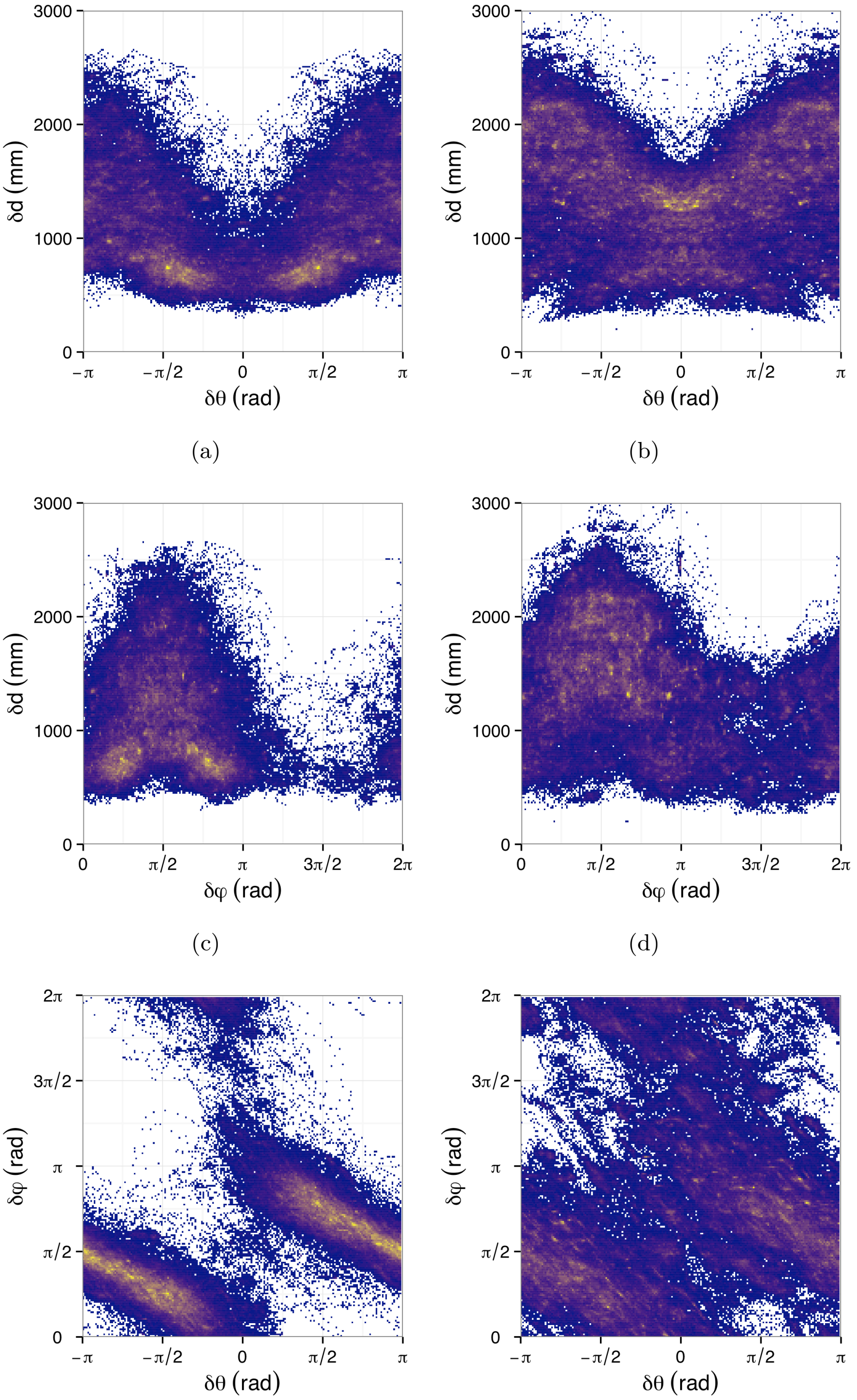
$$\delta\varphi = \operatorname{arctan2}\left(-d'_2, -d'_1\right) + \pi \quad \text{where} \quad d' = \left(R^{IOC}\right) R^{MC}_{j,0} \left(R^{MC}_{j,t}\right)^{T} d \;. \tag{20}$$

Note that the pair of δd and $\delta\varphi$ can be interpreted as magnitude and argument in the domain of complex numbers, which is why $\delta\varphi$ has been defined in $[0, 2\pi)$ (hence the change of signs and the increment about π).

2.2.5 *The final dataset*

The previous sections documented the recording and post-processing of the experimental data. For each frame and pair of persons, the variables $\delta\theta$, δd and $\delta\varphi$ were computed, yielding a pair of observations of interaction geometry as seen from either person. The data were visualized and annotated, thence providing ground-truth for social situations. The results of the annotation were double-checked and verified by thorough analysis of both the visualization and the time-stamped video footage. From here, the dataset can be split into two partitions ($S^{\oplus}$, $S^{\ominus}$), one representing the pair-wise observations of those persons who were engaged in social interaction and would hence be part of a social situation ($S^{\oplus}$), and one for the pairs of persons who would not interact with each other, meaning they were either not part of the same social situation or did not participate in any social situation at a given time ($S^{\ominus}$). Note that the beginning of the first as well as the ending of the last social situation constitute temporal boundaries for *both* partitions, since outside of this interval there is no explicit ground-truth for either $S^{\oplus}$ or $S^{\ominus}$.

The final dataset consists of 368234 observations for $S^{\oplus}$ and 457318 for $S^{\ominus}$, verified against the number of frames as well as the number of distinct pairs $\binom{N}{2}$ for every social situation with arity N. Figure 7 shows bivariate histograms of the observations for $S^{\oplus}$ and $S^{\ominus}$. For $S^{\oplus}$, all three of the pairs ($\delta\theta$, $\delta\varphi$), ($\delta\theta$, δd) and ($\delta\varphi$, δd) feature clearly defined clusters. Note the inherent periodicity of $\delta\theta$ and $\delta\varphi$, from which it follows that the apparent two distinct clusters in the histogram of ($\delta\theta$, $\delta\varphi$) actually form a single cluster instead. Similar to $S^{\oplus}$, several clusters can be found for $S^{\ominus}$, yet their edges are a lot fuzzier and their distributions are generally wider. One may argue that the similarities of $S^{\ominus}$ to $S^{\oplus}$ were a consequence of the experimental settings, for example caused by the constrained area in which people could move, or other potential influences such as personal profile parameters of the participants. In this regard, one may in particular expect $S^{\ominus}$ to follow a complete random white noise distribution. It should be noted, however, that $S^{\ominus}$ will not adopt such a distribution even under an infinite number of observations. As discussed in chapter 1, social behaviour dictates a certain degree of perceived "non-awkwardness" [166], manifested e. g. in the fact that humans strive to either clearly establish or separate from social situations. For example, one person standing close to and in front of another person would usually imply a certain sense of awkwardness, and such behaviour is typically avoided, except for situations where that is simply impossible, for instance when riding an overcrowded subway. Generally speaking, the marginal and joint distributions of the observed variables conform to the elementary and intuitive expectations towards interaction geometry in human behaviour. The following sections discuss the distributions of $\delta\varphi$, $\delta\theta$ and δd in more detail.



(a)

(b)

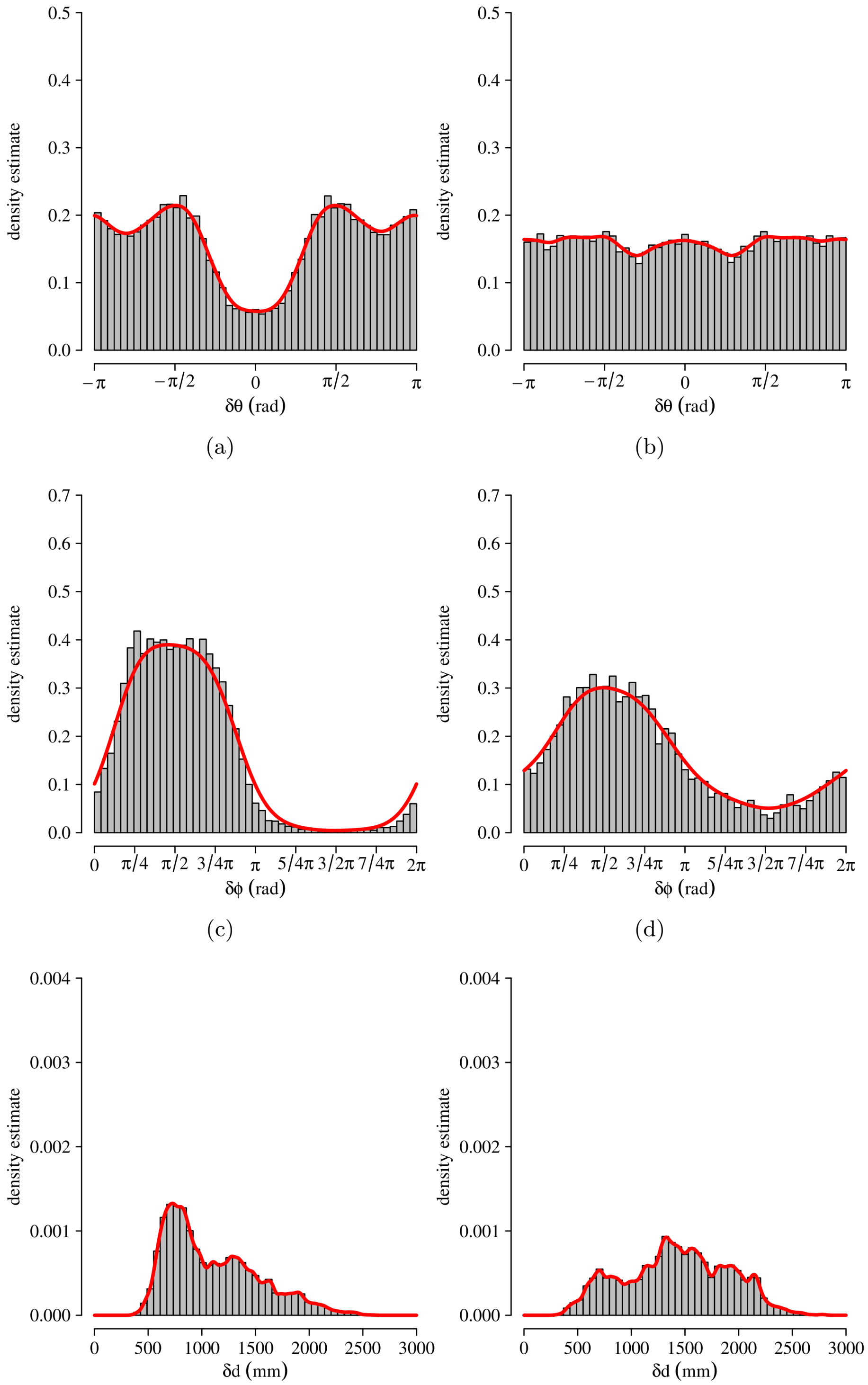
(c)

(d)

(e)

(f)

Figure 7.: Color-coded histograms of the joint distributions of $\delta\theta$, $\delta\varphi$ and δd for classes $S^{\oplus}$ (a,c,e) and $S^{\ominus}$ (b,d,f).



(a)

(b)

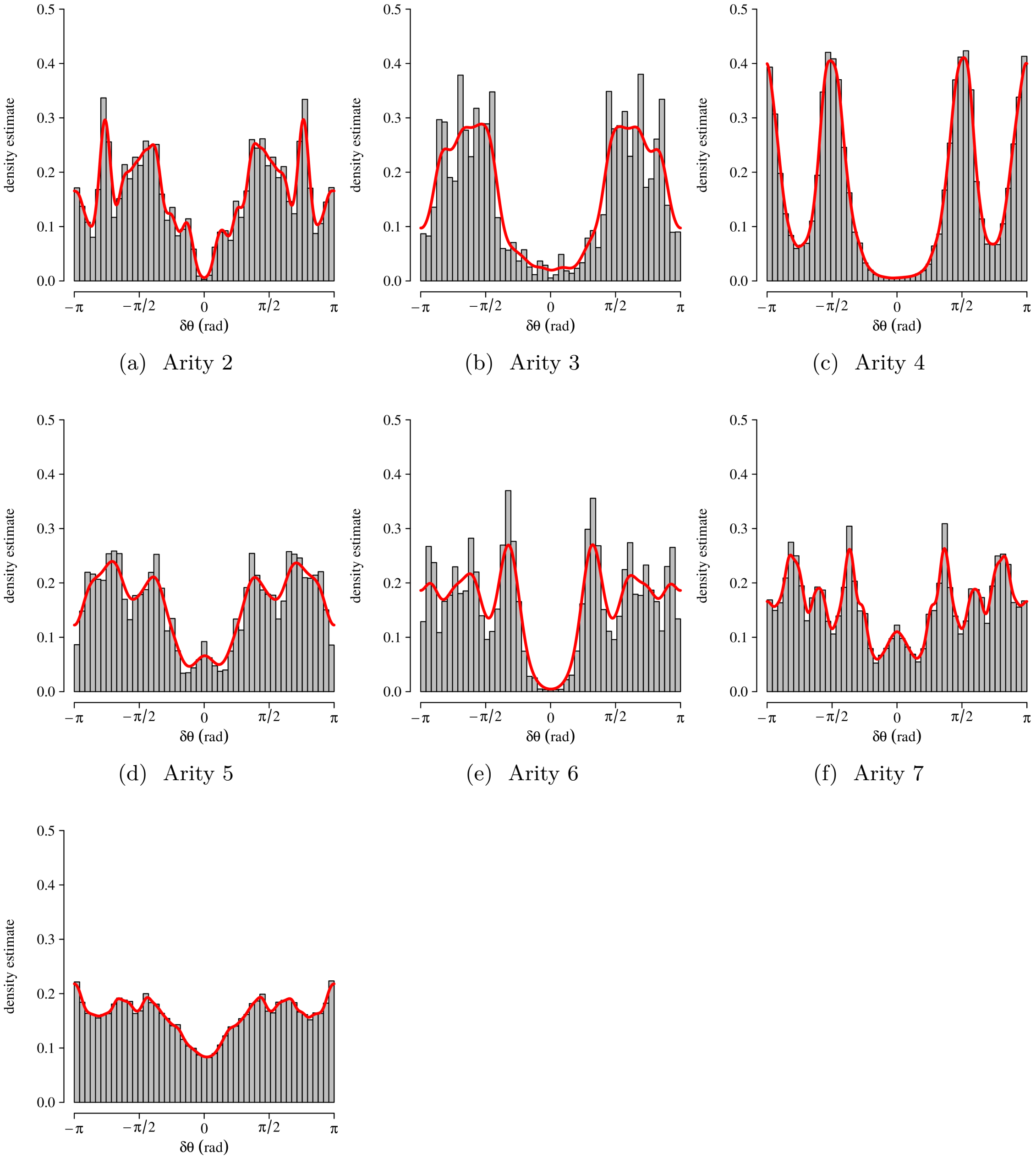
(c)

(d)

(e)

(f)

Figure 8.: Histograms and kernel density estimations of the distributions of $\delta\theta$, $\delta\varphi$ and δd for $S^{\oplus}$ (a,c,e) and $S^{\ominus}$ (b,d,f), using a Gaussian kernel and bandwidths of $10°$, $10°$ and 25 mm for $\delta\theta$, $\delta\varphi$ and δd, respectively.



(a) Arity 2

(b) Arity 3

(c) Arity 4

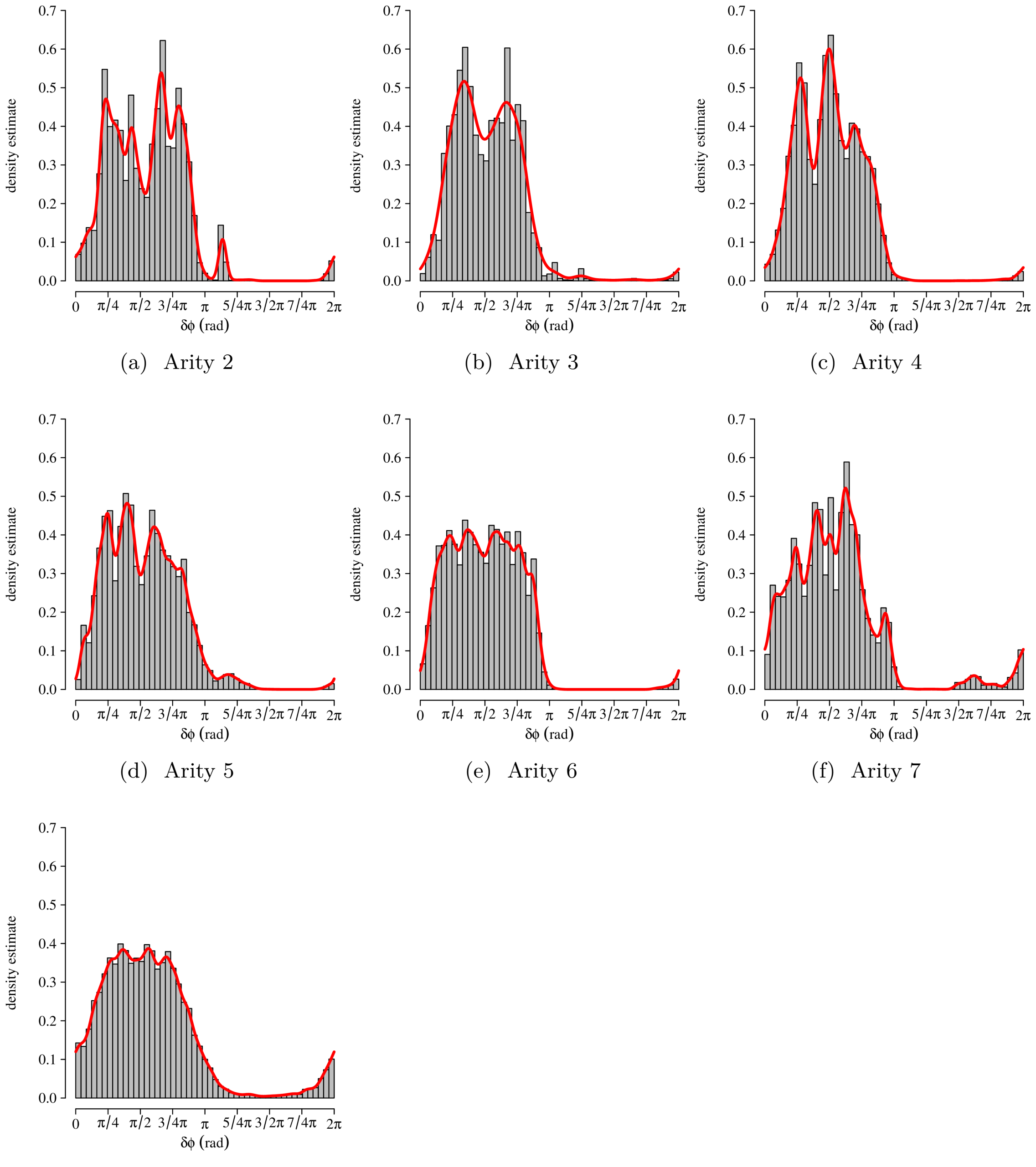
(d) Arity 5

(e) Arity 6

(f) Arity 7

(g) Arity 9

Figure 9.: Histograms and kernel density estimations of δθ for varying arities, using a Gaussian kernel and bandwidths of 5° (a,c,f,g) and 10° (b,d,e).



(a) Arity 2

(b) Arity 3

(c) Arity 4

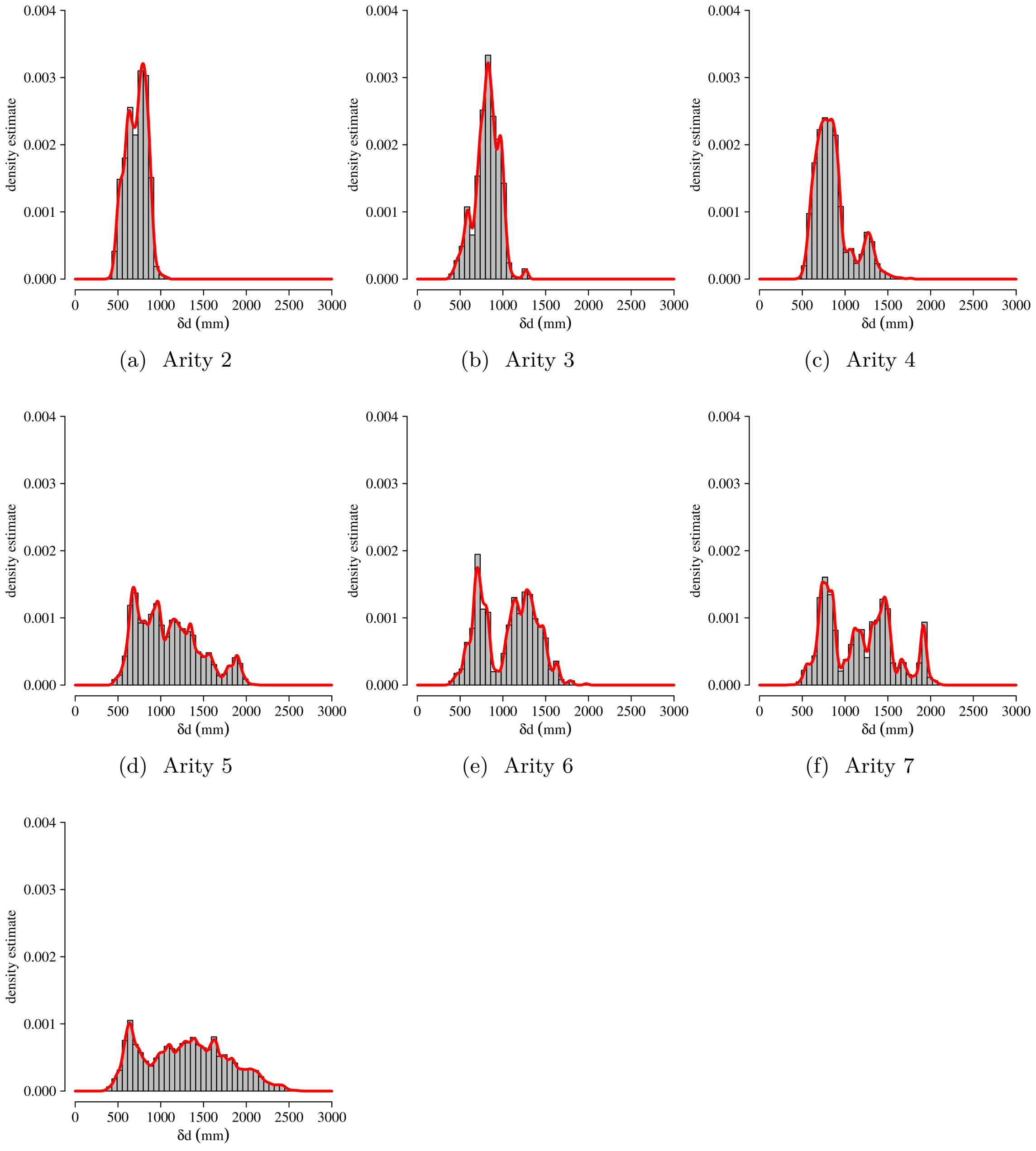
(d) Arity 5

(e) Arity 6

(f) Arity 7

(g) Arity 9

Figure 10.: Histograms and kernel density estimations of $\delta\varphi$ for varying arities, using a Gaussian kernel and bandwidths of 5° (a,c,d,e,f,g) and 10° (b).



(a) Arity 2

(b) Arity 3

(c) Arity 4

(d) Arity 5

(e) Arity 6

(f) Arity 7

(g) Arity 9

Figure 11.: Histograms and kernel density estimations of δd for varying arities, using a Gaussian kernel and a common bandwidth of 25 mm.

2.2.5.1 $\delta\varphi$

According to the distribution of $\delta\varphi$ for $S^{\oplus}$ (figure 8c), interactions take place almost exclusively in front of a person ($\delta\varphi \in [0, \pi)$). There is however a non-negligible number of observations close to 2π in a person's rear hemisphere. This is interesting and can be explained for two reasons: First, people might briefly turn to look at somebody or something else while still maintaining the same social situation. Second, aside from turning, observations of $\delta\varphi$ close to and/or between π and 2π typically occur whenever additional people enter an already established social situation. Imagine, for example, a situation where five people stand in a circular formation and a sixth person signals their wish to enter that situation by approaching the circle from the outside, and possibly even standing there for a short while until the circle finally opens and that sixth person is included. Both explanations are backed by the joint distribution of $\delta\varphi$ and δd (figure 7d), which clearly shows that most back-side observations close to 2π occurred at short distances of about 70 cm. The distributions of $\delta\varphi$ for social situations with 7 or 9 participants (figures 10f and 10g) provide further evidence as in both cases the number of observations close to 2π are significantly higher than for smaller arities.
Furthermore, partitioning the set of observations for $S^{\oplus}$ by group cardinality exhibits typical configurations in social interaction geometry. Figure 10 shows histograms and kernel density plots for $\delta\varphi$ and varying arities. Note that this is only provided for $S^{\oplus}$ because there is no meaningful way to tell whether a person was *not* in a social situation with a *defined* number of others. Here, the correlation between the distribution of $\delta\varphi$ and the arity of the corresponding situation can clearly be seen, and the local maxima of the variable tend to comply with basic expectations. Circular formations are typical [166], especially along with an increasing number of group members, for which one would expect more or less *evenly* distributed positions on a semicircle. For example, the "ideal configuration" (so to speak) of 3 persons would imply that these persons are mutually located at angles of about $60°$ as seen from each individual shoulder-line. This is clearly reflected by the respective distribution of $\delta\varphi$ (figure 10b), where from any person's point of view, other interactants would most often stand at angles of $60°$ and $120°$. The same naturally holds for groups of 4 (figure 10c) or more persons. Table 3 compares the actual local maxima of each distribution of $\delta\varphi$ with the "ideal" configuration per arity. Notably, while $N = 3$ to $N = 7$ fulfill the expectations, this does not seem to be the case for $N = 2$ and $N = 9$. In fact, the variance of the distribution for $N = 9$ is high enough so that there are no obvious local maxima. However, this comes to no surprise as higher cardinalities force smaller gaps and increased distances between the persons' positions on the circle, which is why relatively small movements or rotations of a person already have a huge influence on one or all of the observed variables. Lastly, $N = 2$ is a special case in so far as the theoretical ideal configuration at $90°$ can hardly be observed. In fact, it would imply a strictly frontal pose which is rather found in formal settings like talking to a superior at work [206]. One may further note that, during the experiment, people regularly exhibited a certain openness towards others. This could relate to the basic personal need of moni-

Arity	Local maxima (degrees)	Ideal configuration (degrees)
2	42, 78, 119, 144	special case
3	61, 120	60, 120
4	50, 90, 125	45, 90, 135
5	44, 71, 110, 146	36, 72, 108, 144
6	42, 66, 106, 137, 154	30, 60, 90, 120, 150
7	16, 43, 73, 90, 113, 168	26, 51, 77, 103, 129, 154
9	66, 101, 126	20, 40, 60, 80, 100, 120, 140, 160

Table 3.: Local maxima of $\delta\varphi$ vs. evenly distributed positions on the semicircle.

toring one's environment and/or the persons therein, but also signals preemptive approval or invitation of outsiders to join a social situation.

Apart from $\delta\varphi$ for $S^{\oplus}$, it is striking that the distribution is surprisingly similar for $S^{\ominus}$ (figure 8d). Both distributions actually feature a peak at about $90°$ (front) and a trough around $270°$ (rear). Still, the number of observations in the front is much less and more evenly distributed for $S^{\ominus}$ when compared to $S^{\oplus}$, and while there are no notable observations in the rear for $S^{\oplus}$, there are considerably more for $S^{\ominus}$ within $[\pi, 2\pi)$. The fact that $S^{\ominus}$ does not contain more observations in this particular interval is well worth discussing. Indeed, this seems to be caused by experimental effects, namely the restricted area in which people could freely move and still be recorded by the cameras. During the experiment, in order to face others, the participants would often stand with their backs towards the walls and hence the boundaries of the recording area. The confined space would consequently not allow for others to stand behind their backs. In this regard, the joint distribution of $\delta\varphi$ and δd (as shown in figure 7d) exhibits a lack of observations at distances of more than about 1.5 m. In real life one would rightly expect the number of observations to grow with increasing distance in case of $S^{\ominus}$, which is especially true for the interval $[\pi, 2\pi]$. Moreover, it is legitimate to claim that a higher number of samples, and thus a more (but not completely) even (joint) distribution of $\delta\varphi$ (and δd) for $S^{\ominus}$, would eventually lead to greater differences between both classes, thereby considerably simplifying the work of a classifier. Hence the dataset is considered to be on the "safe side". Aside from this effect, it is interesting that in case of $S^{\ominus}$ there is a noticeable "hole" around $\delta\varphi \approx 4/3\pi$ and $\delta d \approx 1m$, indicating that people tend to establish geometric constellations in a way such that, at close distances, other persons will not be located directly behind but preferably more to the sides of their backs.

Overall, the correlation of $\delta\varphi$ and δd is high. The joint distribution of $\delta\varphi$ and δd for $S^{\oplus}$ (figure 7c) augments the marginal distribution of $\delta\varphi$ in so far as most interactions occur at about $\delta\varphi \approx \pi/3$ and $\delta\varphi \approx 2/3\pi$ at distances of less than 1 m, which means that the vast majority of interactions is more to the side, which in turn supports the theory of perceived openness of social situations towards others. For $S^{\oplus}$, the peaks are also way more pronounced, whereas for $S^{\ominus}$, the distribution clearly lacks distinct local maxima. It is also worth mentioning that, for $S^{\oplus}$, there is a noticable gap between ~ 60° and ~ 90° at particularly short distances of less than 75 cm (see figure 7c). This makes sense because this area is covered by the intimate and personal zones as defined by Hall [133], and neither any constraints of the experiment nor the personal background of the participating persons would allow one person to stand respectively close in front of another person without being at least being perceived as awkward. The gap is even bigger for $S^{\ominus}$, ranging up to ~ $1.25\ m$ (see figure 7d). Naturally, standing with the back towards another person at such distances is rather "unappropriate" and would not follow the common sense of social behavior.

At last, note that the variance of $\delta\varphi$ decreases with increasing δd. This is characteristic because the farther interacting persons are apart, the more likely they attempt to face each other and hence $\delta\varphi$ slowly approaches $\pi/2$. This is verified by the more moderate distributions of $\delta\varphi$ for increasing group sizes (figure 10). It also reflects the constraints that group cardinality imposes on FFS, in particular a tendency towards wider circular formations as more persons participate.

2.2.5.2 $\delta\theta$

In contrast to $\delta\varphi$, the values of $\delta\theta$ are symmetric for any pair of persons, no matter whether they are interacting or not. According to the overall distribution of $\delta\theta$ (figure 8a), most interactions occur at angles of ± 80 and ± 170 degrees between shoulder-lines, and only very few around $0°$, i. e. whenever the shoulder-lines of two persons are aligned in parallel and both are facing the same direction. It should be noted that the extrema at ± 170 degrees are quite close to a full frontal configuration at $180°$, which in turn further sustains the discussed tendency to avoid such poses. Also note the local minima around $140°$ which nonetheless differ from the global minimum at $0°$. Similar to the observations of $\delta\varphi$ close to 2π, the non-negligible number of observations of $\delta\theta$ around $0°$ is a consequence of either people approaching an already established social situation from the outside, specifically in the rear of other participants, or even more likely the observable fact that people tend to turn into the direction of the current speaker or dominating person, where more often than not the shoulder-lines of adjacent persons shift into similar or equal alignment. This effect is particularly noticeable in larger groups, but applies to smaller ones as well. In contrast to the peaked distribution for $S^{\oplus}$, $\delta\theta$ is much more evenly distributed for $S^{\ominus}$ (figure 8b). Interestingly, the minima at ~ ±55 degrees for $S^{\ominus}$ are not strictly opposite to the maxima for $S^{\oplus}$. Still, they represent an orientation which is rather likely to be interpreted as mutual interaction either taking place or being about to be established.

Furthermore, partitioning the dataset by group cardinality confirms typical geometrical configurations during social situations (figure 9). Common to all distributions for varying group sizes is a global minimum at $0°$. The distributions for arities 3 to 7 (figures 9b, 9c, 9d, 9e, 9f) are consistent with the expectations for "ideal configurations" of the respective number of persons (table 4), whereas group sizes of 2 and 9 are somewhat distinct, as was

Arity	Local maxima (degrees)	Ideal configuration (degrees)
2	±49, ±71, ±138, ±172	special case
3	±96, ±118, ±147	±120
4	±93, ±172	±90, 180
5	±70, ±128	±72, ±144
6	±58, ±111, ±163	±60, ±120, 180
7	±66, ±108, ±148	±51, ±103, ±154
9	±78, ±118, ±171	±40, ±80, ±120, ±160

Table 4.: Local maxima of $\delta\theta$ vs. evenly distributed orientations along the semicircle.

the case for the observed vs. ideal configurations in case of $\delta\varphi$ (refer to table 3). Again, the much higher variance (as a consequence of the fact that smaller shifts cause greater changes in the observed variables for larger group sizes) is what leads to the specific shape of the distribution for groups of 9. On the other hand, the distribution for groups of 2 features several spikes due to the higher number of possible (and typical) geometrical configurations, as opposed to larger groups. Still, typical configurations are observable at ±70, ±140 and ±170 degrees. Interestingly enough, variance was the least in groups of 4, implying rather static spatio-orientational formations. This may be surprising, as 5 groups of 4 persons were observed over an accumulated duration of more than 10 minutes throughout the whole experiment (figure 5). The same effect is however not observable for other group cardinalities with comparable duration. In accordance with the prior reasoning, the distributions for arities 5, 7 and 9 show considerably more observations around 0°, providing further evidence for the hypothesis that it is more likely for two adjacent persons in larger groups to shift towards the same orientation and subsequently face the same speaker. This is also corroborated by the video footage of the corresponding groups. From these distributions, one might consequently expect the same for groups of 6, yet the respective distribution lacks one such peak. This is likely caused by the fact that, overall, there were only 2 situations with groups of 6, both of which did not last for more than a couple of minutes in total.

The basic shapes of the joint distributions of $\delta\theta$ and δd look similar for both classes (figures 7a, 7b). Closer investigation of $S^{\oplus}$ reveals peaks for interactions which mostly occur at shoulder-line angles around 80° and distances between 60 cm and 90 cm. This is distinct from $S^{\ominus}$, for which observations are almost evenly distributed among the whole domain except for the two areas between ±80° and ±180° at distances of up to 1 m. The latter is clearly opposite to $S^{\oplus}$ and conforms to intuitive expectations of social behavior. In regard of $S^{\ominus}$, angles around 0° within the same range of interpersonal distances naturally hint at a lack of interaction, which is also expected. Variance is proportional to increasing distance, and the relative frequency of observations is more evenly distributed. Note that both distributions have few to none observations at 0° and distances of 1.75 m ($S^{\ominus}$) respective 1.25 m ($S^{\oplus}$) and above. This is another unfortunate effect of the constrained environment. However, further note that there are way less corresponding observations for $S^{\oplus}$ than for $S^{\ominus}$. This, plus the fact that the distribution for $S^{\oplus}$ has the highest frequency of observations where it is lowest for $S^{\ominus}$, lead to the conclusion that potential influences due to environmental constraints can be considered negligible, similar to the corresponding effect that was observed for the joint distribution of $\delta\varphi$ and δd. One may further note that the bottom shape of the distribution for $S^{\oplus}$ is rather convex while it is concave for $S^{\ominus}$. Full frontal configurations are avoided independent of the presence of social interaction. This is consistent with expectations towards the various possible geometric configurations as well as the implications for larger groups. The joint distribution of $\delta\theta$ and $\delta\varphi$ shows very high linear and non-linear correlation for $S^{\oplus}$, where two clusters clearly emerge, which is not the case for $S^{\ominus}$, for which, again, the data are evenly distributed over large areas. The correlation between $\delta\theta$ and $\delta\varphi$ is indeed meaningful because – during, but not restricted to interaction – mutual orientation naturally depends on the relative position (both angle $\delta\varphi$ and distance δd). For example, one can observe almost full-frontal orientations at likewise frontal positions, and flatter angles to the sides. All the same, the less populated areas in the rear ($3\pi/2\delta\varphi$), together with relative frontal configurations ($\pm\pi\delta\theta$), relate to the same environmental restrictions that were discussed before.

2.2.5.3 δd

The greatest qualitative difference between $S^{\oplus}$ and $S^{\ominus}$ can be seen with respect to interpersonal distance δd (figures 8e and 8f). For $S^{\oplus}$, there is a significant peak at 72 cm, as well as a second, not so pronounced, peak at 128 cm. The former peak corresponds to the personal zone while the latter is located shortly after the beginning of the social zone. There are no observations of δd below 50 cm, i. e. inside the intimate zone. On the other hand, for $S^{\ominus}$ one notices three areas with peaks at about 70 cm, 130 cm and 210 cm. Beyond that, the number of observations decreases with a much greater slope than in case of $S^{\oplus}$. Furthermore, for $S^{\oplus}$ the decrease already starts at or even before 200 cm and is almost monotonic. For both classes, the number of observations vanishes almost completely at about 250 cm. This is, at least in case of $S^{\ominus}$, certainly a consequence of the confined space during the experiment. In general, one would naturally expect a continuous growth of the number of observations along with increasing distance, and likewise the contrary for $S^{\oplus}$, for which the range of the intimate zone would impose a subtle but clear constraint. Interestingly enough, according to figure 8f, these expectations are not generally met for $S^{\ominus}$, at least not within a range of up to 250 cm. It is however clear that in spite of the fact that interactions do also occur at greater distances, e. g. within the public zone, a threshold could be selected after which the general probability of social interaction is less than for no interactions at all, and hence a classifier could decide for $S^{\ominus}$ whenever this threshold were to be exceeded. The present dataset does not allow for such a threshold to be well selected.

The explicitness of the peak at 72 cm for social interactions is merely a consequence of the fact that all distributions exhibit a similar peak even when the data are split according

Arity	Local extrema (cm)	Ideal configuration (cm)
2	63, 79	70
3	59, 83	70
4	80, 127	70, 99
5	68, 96, 116, 135, 157, 180	70, 113
6	70, 115, 128, 146, 163	70, 121, 140
7	74, 120, 146, 192	70, 126, 157
9	64, 110, 138, 162, 205	70, 132, 177, 202

Table 5.: Local extrema vs. ideal distances assuming 70 cm between adjacent persons in circular formation.

to group size (figure 11). From this, it follows that this specific value for interpersonal distance is the most representative for the personal zone, and potentially also perceived as comfortable and socially best acceptable, given the circumstances of the experiment. It goes without saying that this relates to adjacent persons only, and therefore that it is basically independent of the cardinality and geometrical configuration of a group. As a consequence, it is possible to define the ideal mutual distance between any member of a group in circular configuration. Given this circular shape constraint, plus the constraint that neighbours should be located 70 cm apart from each other, the ideal mutual distance can thus be determined per group size. Table 5 compares these theoretical distances to the local extrema of the actual distributions. For this, recall that δd is measured between the center of the bodies, and not between adjacent shoulders or the shortest distance between any respective body parts.

On a final note, the distributions of δd for arities of 2 and 3 are unimodal and exhibit relatively small variance. Among all, groups of 3 feature the most distinct peak in comparison to the average of 72 cm. For groups of 4, apart from a peak at about 80 cm, yet with greater variance, one also notices an increased number of observations at 128 cm. Assuming circular configuration and ideal distances of 70 cm between adjacent persons, both values of 70 cm and 98 cm are well explained by the first peak. Manual analysis of the corresponding video footage and visualizations unveils that the second peak is in fact caused by two particular members of a group of 4 who stood farther apart from each other for a quite long period of time. In addition to that, another member of this group occasionally walked back and forth a few steps, hence causing greater variance than commonly found in equally sized or greater groups. Arguably except for groups of 6, groups of 5 or more participants feature a much more equal distribution of δd. Note that groups of 6 and 7 occurred only a few times and only during short periods, hence the relatively small amount of data is less meaningful for these than for the rest. According to the video footage, larger groups most of the time established approximate circular formations, yet all the same their formations differed from ideal and static circular formations every now and then. The latter mostly occurred when the dynamics of other groups or individuals forced members of a particular group to move accordingly.

2.2.5.4 *Discussion*

The previous analysis of the marginal and joint distributions of the $\delta\theta$, $\delta\varphi$ and δd supports both the applicability and expressiveness of the dataset for its use in interaction geometry. Most notably, fundamental expectations towards spatio-orientational behaviour are satisfied and well-reflected in the recorded data. Eventually, the analysis shows that interaction geometry can indeed lead to well interpretable and manageable models. The validity of the data is further corroborated by statistical analysis of the correlations between the variables. Table 6 shows the correlation matrices where each element corresponds to the Spearman correlation coefficient ρ, which, contrary to the Pearson coefficient, can also express non-linear relations. For this, $\rho \in [-1, +1]$ denotes whether one variable can be described by another variable through some monotonic function.
Analysis of the full dataset for $S^{\oplus}$ shows a strong correlation between relative position $\delta\varphi$ and relative orientation $\delta\theta$, whereas the same relation is much less for $S^{\ominus}$ (tables 6a 6b). Perhaps surprisingly, the correlation between $\delta\theta$ and relative distance δd is close to none for both classes. Even more so, it appears as if the correlation between $\delta\varphi$ and δd were much stronger for $S^{\ominus}$ than for $S^{\oplus}$, which in turn would contradict the discussed expectations towards proxemic behaviour. It turns out that the apparent problem is in fact rooted in the symmetry of $\delta\theta$ and δd. Recall that, for each pair of persons and any particular time frame, δd_{ij} is equal to δd_{ji}, and $\delta\theta_{ij}$ and $\delta\theta_{ji}$ differ in sign, but not in magnitude (except for random measurement errors). Furthermore, note that $\delta\varphi$ depends on $\delta\theta$ to a large degree, so that the symmetry of $\delta\theta$ is again responsible for the low correlation coefficient for $\delta\varphi$ and δd in case of $S^{\oplus}$. The present issue can be easily resolved by considering only one out of two corresponding samples for each time frame, thus effectively reducing the data to half size (see below for a further discussion of symmetry). For the adapted dataset, the strong relation between $\delta\theta$ and $\delta\varphi$ is emphasized by Spearman's correlation coefficient even more. It furthermore exhibits a significant correlation between $\delta\theta$ and δd, as well as a less strong, but still noticeable, relation between $\delta\varphi$ and δd. The discrepancies between the correlations of $\delta\theta$ and $\delta\varphi$ for $S^{\oplus}$ and $S^{\ominus}$ are striking, and are much less in case of the other variables. This is again presumably an effect of the constrained recording area during the experiment. Still, the correlations of $\delta\theta$ and $\delta\varphi$ with δd are relatively higher for $S^{\oplus}$ than for $S^{\ominus}$. In case of $S^{\ominus}$, it may be expected that any δd-related coefficients will tend to zero once more and more data were collected, particularly so in unconstrained environments. At last, the apparent contradiction that the correlation coefficient between $\delta\varphi$ and δd is higher for $S^{\ominus}$ than for $S^{\oplus}$ is also explained through the latter analysis of the reduced dataset. Table 6c reveals a noticeable correlation between $\delta\varphi$ and δd when compared with table 6a. It should be noted that the reduction of the dataset had no influence on the corresponding value of ρ when comparing tables 6d and 6b. This shows that the present issue is indeed explained by the (expected) noisy nature of the data for $S^{\ominus}$.

	$\delta\theta$	$\delta\varphi$	δd
$\delta\theta$	1.000	0.481	-0.003
$\delta\varphi$	0.481	1.000	-0.072
δd	-0.003	-0.072	1.000

(a) $S^{\oplus}$ full

	$\delta\theta$	$\delta\varphi$	δd
$\delta\theta$	1.000	0.233	-0.006
$\delta\varphi$	0.233	1.000	-0.328
δd	-0.006	-0.328	1.000

(b) $S^{\ominus}$ full

	$\delta\theta$	$\delta\varphi$	δd
$\delta\theta$	1.000	-0.750	0.556
$\delta\varphi$	-0.750	1.000	-0.440
δd	0.556	-0.440	1.000

(c) $S^{\oplus}$ reduced

	$\delta\theta$	$\delta\varphi$	δd
$\delta\theta$	1.000	-0.372	0.438
$\delta\varphi$	-0.372	1.000	-0.345
δd	0.438	-0.345	1.000

(d) $S^{\ominus}$ reduced

Table 6.: Spearman correlation coefficients for the final dataset

As discussed before, the variables' distributions (refer to figure 7) suggest inherent symmetries in a part of the data. This is obviously the case for $\delta\theta$ and δd, but $\delta\varphi$ is not strictly symmetrical. It is however possible to define a function $f : [-\pi, +\pi] \times [0, 2\pi] \to [0, 2\pi]$ where

$$f : (\delta\theta_A, \delta\varphi_A) \mapsto \delta\varphi_B = \pi + \delta\varphi_A + \delta\theta_A \tag{21}$$

which allows for computing $\delta\varphi_B$ from the samples measured by A. Hence the data for A and B are symmetrical in the sense that all variables can be determined for both persons based solely on the measurements of either person, irrespective of whether A and B are members of the same social situation. Depending on the mathematical model to be used, let alone the size of the final dataset with 368,234 + 457,318 = 825,552 samples in total, any apparent symmetry could perhaps be used to one's advantage, for example by reducing the amount of data for an improved memory footprint and less computational cost. Moreover, one should consider the degree to which the redundancy of symmetrical data might be disadvantageous for a potential classifier. Very generally speaking, this depends on both the chosen classifier as well as any specific kind of redundancy. For the present dataset, however, it merely corresponds to partial mirroring, and will have no negative impact on the model and classifier as discussed in the forthcoming sections. More importantly, removal of redundant symmetries would need to be explicitly enforced for any number of newly gathered observations. So, for any given pair of observations, selecting one over the other must be subject to predefined criteria and would be imply further processing. Most notably, reducing the dataset would first and foremost propagate or even increase any systematic and/or random measurement errors. Indeed, actually cutting the present dataset in half, and subsequently computing either half based on the other, yields a noticeable error for both $\delta\theta$ and $\delta\varphi$, as can be seen from table 7. In order to avoid the introduction of additional random measurement errors the dataset was therefore considered as a whole. Due to the fact that the pairwise symmetries occur for both $S^{\oplus}$ and $S^{\ominus}$ at once, it is ensured that leaving the dataset as-is will not lead to overfitting or otherwise adverse effects.

	$\delta\theta$	$\delta\varphi$
$S^{\oplus}$	$3.24\ deg^2$	$4.73\ deg^2$
$S^{\ominus}$	$4.62\ deg^2$	$4.81\ deg^2$

Table 7.: Mean squared error upon removal of presumed redundancies.

2.3 MODELS FOR INTERACTION GEOMETRY

The following sections develop an appropriate mathematical model for automatic detection of social interaction based on interaction geometry. Ideally, such a model would allow for thorough understanding and easy interpretability, in particular with respect to socio-psychological research. It will be shown that interaction geometry allows for probabilistic decisions upon the presence ($S^{\oplus}$) or absence ($S^{\ominus}$) of *dyadic* social interaction for any pair of persons (i, j) at any time t. Social situations of greater cardinalities can then be inferred e. g. by means of graph clustering. Note that analysis and interpretation of group phenomena based on dyads is common practice in social sciences [231, 84, 67, 120, 205].

The proposed model is a *generative* model deduced from the accumulated $(\delta\theta, \delta\varphi, \delta d)_{ij}$ over all time frames and ordered pairs $\{(i, j) \mid i, j \in P, i \neq j\}$, where P denotes the set of persons in the data. For the present task, generative models are preferable over *discriminative* models. Discriminative models arguably have the advantage of deciding a classification problem without the need for explicitly modeling the probability densities of the features [218]. They allow for almost arbitrary preprocessing of the features, such as the application of kernel functions prior to fitting the model, and they are supposed to exhibit better performance than generative models on discrete tasks [34, 218]. This however automatically implies that continuous variables would have to be discretized first, which may lead to an enormous increase in model parameters, especially so for multidimensional data for which the corresponding increase would obviously be exponential, a fact well-known as the *curse of dimensionality* [34]. This would likewise require a much greater set of training data. Moreover, discriminative models can be considered “sub-symbolic” in the sense that they are typically intractable in terms of interpretability and traceability, such as e. g. the warped decision surfaces of high-dimensional Support Vector Machines (SVMs). Generative models, on the other hand, can be understood as Bayesian Networks and are thus particularly well-suited for those tasks. In spite of the fact that they require a potentially more complex modeling of the observed variables’ probability densities, their advantages for the present task are accounted for as follows: For a given training dataset of N samples, generative models maximize the joint log-likelihood

$$\sum_{i=1}^{N} \log p(x_i, y_i|\theta) \tag{22}$$

of x_i the observed samples and y_i the corresponding class labels (and possibly additional latent variables), given θ the set of model parameters [34, 218]. The probability term in equation (22) is typically computed from the conditional probabilities $p(x_i|y_i, \theta)$ and the class priors $p(y_i|\theta)$, the latter of which are either modeled according to the classes' relative frequencies, or as fully parametrized probability distributions [218]. Including the class priors in the computation of the posterior distribution is a notable advantage of generative models. As such, class priors help to compensate for unevenly distributed classes in the training data, as shown by application of Bayes' rule

$$p(y_i|x_i, \theta) = \frac{p(x_i|y_i, \theta) \cdot p(y_i)}{p(x_i)} . \tag{23}$$

From equation (23) it furthermore follows that, for a given observation x_i, two classes $y_i = 1$ and $y_i = 2$ can easily be discriminated by selecting the one with the higher posterior:

$$\begin{aligned} p(y_i = 1|x_i, \theta) &\overset{?}{>} p(y_i = 2|x_i, \theta) && (24)\\ \Leftrightarrow \frac{p(x_i, \theta|y_i = 1) \cdot p(y_i = 1)}{p(x_i, \theta)} &\overset{?}{>} \frac{p(x_i, \theta|y_i = 2) \cdot p(y_i = 2)}{p(x_i, \theta)} && (25)\\ \Leftrightarrow p(x_i, \theta|y_i = 1) \cdot p(y_i = 1) &\overset{?}{>} p(x_i, \theta|y_i = 2) \cdot p(y_i = 2) && (26) \end{aligned}$$

Moreover, generative models can be used to generate samples by drawing from $p(y_i|\theta)$ and $p(x_i|y_i, \theta)$ for corresponding y_i. Generative models can hence cope with missing data or, as e. g. in case of Hidden Markov Models (HMMs), input sequences of variable length, and may furthermore aid in the detection of outliers through the marginal $p(x_i)$. For SSP, generative models could otherwise prove useful e. g. for simulating large-scale social situation data. Eventually, existing generative models are much easier to adapt than models based on e. g. non-linear optimization, possibly in real-time and/or on mobile hardware.

2.3.1 *Gaussian Mixture Models*

Figure 7 on page 35 reveals a number of overlaps between the (joint) distributions of $\delta\theta$, $\delta\varphi$ and δd between $S^{\oplus}$ and $S^{\ominus}$. Also, the data in $S^{\oplus}$ appear in significant clusters which are qualitatively easy to distinguish from those in $S^{\ominus}$. This suggests the use of one probabilistic model per class, each based on a *multimodal* distribution. Multimodal distributions, also known as mixture distributions [209], are commonly determined as the superposition of several unimodal distributions. One such distribution is known as Gaussian Mixture Model (GMM), defined as

$$p(x|\theta) = \sum_k \pi_k \mathcal{N}(x|\mu_k, \Sigma_k) \; , \tag{27}$$

subject to

$$0 \leqslant \pi_k \leqslant 1 \quad \text{and} \quad \sum_k \pi_k = 1 \; . \tag{28}$$

Note that the mixing coefficients can themselves be regarded as the probability of each mixture component for explaining a given observation x. GMMs are widely used in data mining, pattern recognition, machine learning, and statistical analysis [34]. Next to classification, typical use-cases are data generation, including completion of missing data [32, 71], and soft-clustering for which distance metrics are modeled as probabilities. It has been shown that GMMs can approximate every continuous density with arbitrary accuracy [218, 34], which makes them an ideal choice for soft-clustering and discrimination when using multiple models along with Bayesian classification. As they are built on top of the well-studied normal distribution, it is quite easy to avoid overfitting, which is obviously required for every classifier, but even more so for applications in proxemics and social sciences.

GMMs belong to the class of Latent Variable Models (LVMs) [218, 34], which assume that the *observed* data correspond to one or more *latent* variables which cannot be directly observed and are hence considered as hidden. LVMs usually require less parameters than other models. As such, latent variables can be regarded as data in compressed form [218]. Since a single D-variate Gaussian has $D + \frac{D(D+1)}{2}$ free parameters, it follows that for GMMs with K components the corresponding count is $K + K\frac{D(D+1)}{2}$, also accounting for the mixing coefficients. GMMs can be further simplified, e. g. by assuming uniformly distributed mixture coefficients, or by adding arbitrary constraints on the shape of the covariances. The downside of models subject to incomplete data or involving latent variables is that model estimation is often difficult, as is the case for GMMs. Aside from using gradient-based or Newton methods [351] for estimation, the Expectation Maximization (EM) algorithm facilitates the learning process and guarantees monotonic convergence, i. e. the likelihood of the model will increase or at least remain constant at during iteration. Nevertheless, as the function which should be optimized is typically not convex, e. g. due to the fact that there are exactly $K!$ equivalent ways for distributing K sets of parameters among a mixture of K components [34], the algorithm will probably converge to a local rather than the global optimum. Other than that, the EM algorithm alleviates the inclusion of potential constraints [218], such as on the distribution of the mixing coefficients in equation (28) or the covariances. Apart from potential reductions of computational overhead, the latter could be exploited to insert domain-specific knowledge into the process. In the context of models for interaction geometry, such constraints could for instance correspond to previous findings from social sciences.

2.3.1.1 *The Expectation Maximization Algorithm*

The EM algorithm allows for maximum likelihood estimation of the parameter set of a model where the training data suffer from missing values or where optimization of the likelihood is analytically intractable, but can be simplified by assuming the existence of missing latent values [32, 34, 218]. As it will be key to both sections 2.3.1.2 and 2.3.2.4, the general idea of the algorithm is first outlined in this section.

In order to illustrate the difficulties when optimizing maximum likelihood for LVMs, let θ denote the full parameter set of such a model. Let $\mathbf{X}$ denote a set of N independent and identically distributed (i.i.d.) observations, and let $\mathbf{Z}$ be a set of N i.i.d. samples from a hidden variable, such that $\forall i : z_i$ corresponds to x_i. Then, given the joint distribution $p(\mathbf{x}, \mathbf{z}|\theta)$, application of the *sum rule* yields the marginal marginal density over $\mathbf{x}$:

$$p(\mathbf{x}|\theta) = \sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{z}|\theta) \tag{29}$$

Since all the $\mathbf{x}_i$ are independent, the log-likelihood of θ given $\mathbf{X}$ is

$$\begin{aligned} \ln \mathcal{L}(\theta|\mathbf{X}) &= \ln \prod_{\mathbf{x}} p(\mathbf{x}|\theta) \\ &= \ln \prod_{\mathbf{x}} \sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{z}|\theta) \\ &= \sum_{\mathbf{x}} \ln \sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{z}|\theta) \; . \end{aligned} \tag{30}$$

As logarithm and sum cannot be exchanged, this function is typically hard to optimize, and in general no closed form solution can be found for its differential [218]. EM circumvents this problem by introducing the so-called *complete data log-likelihood*

$$\ln \mathcal{L}_c(\theta|\mathbf{X}, \mathbf{Z}) = \sum_{i=1}^{N} \ln p(\mathbf{x}_i, \mathbf{z}_i|\theta) \tag{31}$$

assuming that $\mathbf{X}$ and $\mathbf{Z}$ were both observable. It is then possible to reason about the complete data through the *expected value* under the hidden variable's posterior [34, 218]

$$p(\mathbf{z}|\mathbf{x}, \theta) \propto p(\mathbf{x}|\mathbf{z}, \theta) p(\mathbf{z}|\theta) \; , \tag{32}$$

so that the expected value can be defined as a function of θ at iterations t and $t-1$:

$$\mathcal{Q}(\theta, \theta^{t-1}) = \mathbb{E}_{\mathbf{Z}|\mathbf{X},\theta} \left[\ln \mathcal{L}_c(\theta|\mathbf{X}, \mathbf{Z})\right] \tag{33}$$

$$= \sum_{\mathbf{z}} p(\mathbf{z}|\mathbf{X}, \theta^{t-1}) \cdot \ln \mathcal{L}_c(\theta|\mathbf{X}, \mathbf{Z}) \; . \tag{34}$$

This way, the data are first "completed" by estimation of the latent variables' values (E-step), followed by the optimization of $\mathcal{Q}$ with respect to θ (M-step):

$$\theta^t = \mathrm{argmax}_\theta \; \mathcal{Q}(\theta, \theta^{t-1}) \tag{35}$$

The EM algorithm alternates between the E- and M-steps until convergence of either the model parameters or the log-likelihood.

2.3.1.2 *Learning Gaussian Mixture Models*

The adaption of EM to GMMs with K components is straight-forward. Recall the density

$$p(\mathbf{x}|\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \,, \tag{36}$$

and let $\boldsymbol{z}$ be a K-dimensional latent variable with 1-of-K coding, i. e. subject to $z_k \in \{0, 1\}$ and $\sum_k z_k = 1$. Also recall that the mixing coefficients π_k can be regarded as discrete probabilities of choosing the k-th component. Hence define the marginal of $\boldsymbol{z}$ given $\boldsymbol{\theta}$ as

$$p(z_k = 1|\boldsymbol{\theta}) = \pi_k \tag{37}$$

which, due to the 1-of-K coding of $\boldsymbol{z}$, is equivalent to

$$p(\boldsymbol{z}|\boldsymbol{\theta}) = \prod_{k=1}^{K} \pi_k^{z_k} \,. \tag{38}$$

The distribution of $\mathbf{x}$, provided that $\mathbf{x}$ was drawn from the k-th component, can likewise be written as

$$p(\mathbf{x}|z_k = 1, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_k} \,, \tag{39}$$

so that the joint distribution of $\mathbf{X}$ and $\mathbf{Z}$ is eventually given by

$$p(\mathbf{x}, \boldsymbol{z}|\boldsymbol{\theta}) = p(\mathbf{x}|\boldsymbol{z}, \boldsymbol{\theta})p(\boldsymbol{z}|\boldsymbol{\theta}) = \prod_{k=1}^{K} \left(\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right)^{z_k} \,, \tag{40}$$

from which application of Bayes' theorem leads to the posterior

$$p(z_k = 1|\mathbf{x}, \boldsymbol{\theta}) = \frac{p(z_k = 1|\boldsymbol{\theta})p(\mathbf{x}|z_k = 1, \boldsymbol{\theta})}{p(\mathbf{x}|\boldsymbol{\theta})} = \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{l=1}^{K} \pi_l \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l) \,,} \tag{41}$$

known as the *responsibility* $\gamma(z_{nk})$ of the k-th mixture component for the explanation of a given observation $\mathbf{x}_n$. The expected complete data log-likelihood under the posterior of $\boldsymbol{z}$ is therefore given by

$$\begin{aligned}
\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}_{t-1}) &= \mathbb{E}_{\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}}\left[\sum_{n=1}^{N} \ln p(\mathbf{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta})\right] \\
&= \sum_{n=1}^{N} \mathbb{E}\left[\ln\left\{\prod_{k=1}^{K} (\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_{nk}}\right\}\right] \\
&= \sum_{n=1}^{N} \mathbb{E}\left[\sum_{k=1}^{K} z_{nk} \ln\{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\}\right] \\
&= \sum_{n=1}^{N} \sum_{k=1}^{K} \underbrace{\mathbb{E}\left[z_{nk}\right]}_{\gamma(z_{nk})} \ln\{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\}
\end{aligned} \tag{42}$$

The complete data log-likelihood is easily maximized through its partial derivatives for each model parameter in $\boldsymbol{\theta}$. At first, the responsibilities π_k are optimized using a Lagrange multiplier to enforce the constraint $\sum_k \pi_k = 1$, such that

$$\begin{aligned} & \tfrac{\delta}{\delta\pi_k}\left[\ln p(x,z|\boldsymbol{\theta})\right] + \lambda\left(\textstyle\sum_k \pi_k - 1\right) && \overset{!}{=} 0 \\ \Leftrightarrow\quad & \tfrac{\delta}{\delta\pi_k}\left[\textstyle\sum_n \sum_k z_{nk} \ln \pi_k \mathcal{N}(\mathbf{x_n}|\mu_k, \Sigma_k) + \lambda\left(\sum_k \pi_k - 1\right)\right] && \overset{!}{=} 0 \\ \Leftrightarrow\quad & \textstyle\sum_n \frac{z_{nk}}{\pi_k} + \lambda && \overset{!}{=} 0\,, \end{aligned} \tag{43}$$

for which multiplication by π_k and summation over k yields $\lambda = -N$. Using this result in the partial derivative of $\mathcal{Q}$ with respect to π_k then yields the update rule

$$\pi_k^{t+1} = \frac{1}{N}\sum_{n=1}^{N} \gamma(z_{nk})\,. \tag{44}$$

Accordingly, the update rules for $\boldsymbol{\mu}_\mathbf{k}$ as well as $\boldsymbol{\Sigma}_\mathbf{k}$ are given by

$$\boldsymbol{\mu}_\mathbf{k}^{\mathbf{t+1}} = \frac{1}{\sum_n \gamma(z_{nk})}\sum_n \gamma(z_{nk})\mathbf{x_n} \qquad \text{and} \tag{45}$$

$$\boldsymbol{\Sigma}_\mathbf{k}^{\mathbf{t+1}} = \frac{1}{\sum_n \gamma(z_{nk})}\sum_n \gamma(z_{nk})(\mathbf{x_n} - \boldsymbol{\mu}_\mathbf{k}^{\mathbf{t+1}})(\mathbf{x_n} - \boldsymbol{\mu}_\mathbf{k}^{\mathbf{t+1}})^T\,. \tag{46}$$

Iterative computation of the expected responsibilities (E-step) and subsequent maximization of the log-likelihood through adaption of the model parameters (M-step) are repeated until convergence of either the model's log-likelihood or its parameters.

2.3.2 *Semi-Wrapped Gaussian Mixture Models*

Strictly speaking, GMMs represent probability densities over *linear* variables from $-\infty$ to $+\infty$. Two of the variables, $\delta\theta \in [-\pi, +\pi)$ and $\delta\varphi \in [0, 2\pi)$ are however periodic, raising the question whether GMMs constitute a legitimate choice for this particular dataset.

2.3.2.1 *Periodic Variables and Circular Statistics*

Probability distributions over linear variables are usually considered unfit for periodic variables [34], for instance due to the fact that they fail to represent the basic characteristics of circular data. This is easily demonstrated by considering the two samples $\{\frac{1}{4}\pi, \frac{7}{4}\pi\}$ from a 2π-periodic variable. Averaging the samples yields a maximum likelihood estimate of the mean at π, whereas the true mean is obviously located at 0, i. e. exactly opposite. This is illustrated in figure 12. This problem can e. g. be solved by transforming periodic variables such that every value maps to a two-dimensional vector from the origin to a point on the

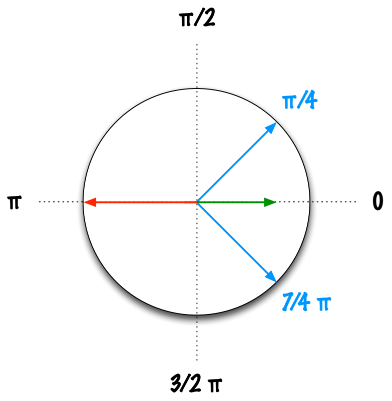


Figure 12.: Circular vs. arithmetic mean. The green vector represents the true circular mean, the red vector the result of averaging the two given samples.

unit circle. These vectors can then be averaged, and the angle between the mean vector and the abscissa determines the true *circular mean*, typically inside the unit circle:

$$\mu_{circular} = \tan^{-1}\left\{\frac{1}{N}\sum_{n}\begin{pmatrix}\cos\,\alpha_n\\ \sin\,\alpha_n\end{pmatrix}\right\} = \arg\left\{\frac{1}{N}\sum_{n} e^{i\alpha_n}\right\} \tag{47}$$

Likewise, the *circular variance* is defined as

$$\nu = 1-\rho = 1-\left\|\frac{1}{N}\sum_{n} e^{i\alpha_n}\right\|\ , \tag{48}$$

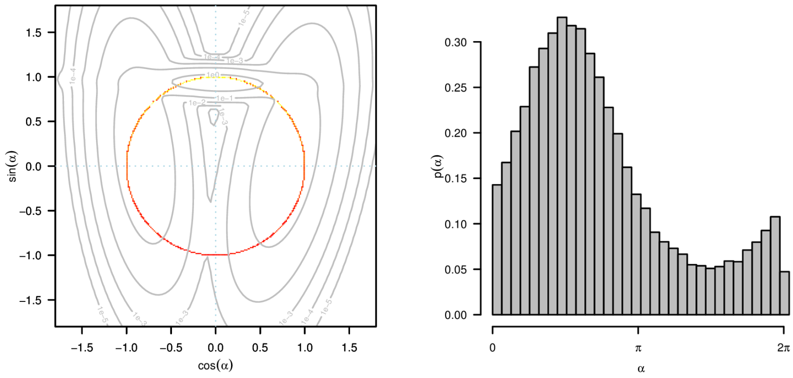
with $0 \leqslant \nu \leqslant 1$. Contrary to the linear case, the *circular standard deviation* is not defined as the square root of ν, but instead as

$$\sigma_{circular} = \sqrt{\ln\frac{1}{(1-\nu)^2}} = \sqrt{\ln\frac{1}{\rho^2}} = \sqrt{-2\ln\rho}\ . \tag{49}$$

This particular form actually turns out to be very useful as an estimate for linear distributions which have been wrapped around the unit circle (see section 2.3.2.2). Circular mean, variance and standard deviation are clearly invariant under rotation, a mandatory property for measures on circular data [200, 34]. In the context of machine learning, rotation invariance is indeed important whenever data need to be *whitened* prior to model training, e. g. through Singular Value Decomposition (SVD), as is the case for the present dataset (see section 2.3.3.1). Comparison of the linear and circular measures for $S^{\oplus}$ and $S^{\ominus}$ indeed yields significant differences for both $\delta\theta$ and $\delta\varphi$ (as shown in table 8). These differences suggest the evaluation of further, potentially more appropriate, models for the probability densities of the present dataset. As a matter of fact, for GMMs, simply projecting the circular data onto a two-dimensional plane is insufficient as it does not change the fact that the respective variables are inherently one-dimensional (see figure 13).

	$S^{\oplus}$				$S^{\ominus}$			
	$\mu_{\mathtt{lin}}$	$\mu_{\mathtt{circ}}$	$\sigma_{\mathtt{lin}}$	$\sigma_{\mathtt{circ}}$	$\mu_{\mathtt{lin}}$	$\mu_{\mathtt{circ}}$	$\sigma_{\mathtt{lin}}$	$\sigma_{\mathtt{circ}}$
$\delta\theta$	~1°	~180°	~113°	~107°	~0°	~180°	~105°	~159°
$\delta\varphi$	~98°	~90°	~60°	~49°	~135°	~94°	~90°	~76°

Table 8.: Comparison of the linear and circular means and standard deviations of $\delta\theta$ and $\delta\varphi$.



(a)

(b)

Figure 13.: (a) Two-dimensional Gaussian Mixture Model on angular data which were previously projected onto the unit circle. (b) Histogram of the true distribution.

2.3.2.2 *Distributions over Periodic Variables*

A number of specific probability distributions exist for periodic variables, starting from basic distributions on the unit circle, like the uniform distribution, the *von Mises* distribution, also known as the “circular normal”, towards more complex ones like the *bivariate von Mises* distribution on the torus, the *Kent* distribution on the two-dimensional unit sphere, or the more general *von Mises-Fisher* distribution on hyper spheres. All of the above are unimodal, and therefore likely the best fit for symmetric data [96]. Multimodality can naturally be accomplished by mixtures of periodic densities [34]. Typical applications for periodic distributions include the analysis of temporal, geological, marine, or metereological data, directional features in handwriting recognition, or the segmentation of color images [96, 21, 180, 275]. Periodic random variables therefore fall into two groups [200], one on which wraps one-dimensional samples around a circle, whereas the other radially projects samples from the two-dimensional plane onto the unit circle, e. g. corresponding to angles. Whereas periodic distributions may excel in terms of statistical properties, one major drawback is that their analytical forms are likely difficult to handle, e. g. when forming joint distributions from multiple periodic and/or linear variables [21, 275]. Likewise, their computational evaluation tends to be expensive [96], such as e. g. in the case of the von Mises distribution whose integral has to be evaluated numerically [96].

Another way for dealing with circular data is through non-parametric methods like histograms [34] or kernel density estimators [183]. The former are flexible and easy to handle, but suffer from limitations such as the optimal choice for the width of the bins, with small bins tending to spiky and huge bins tending to overly smoothed distributions. Furthermore, both are prone to quantization errors, and may consume a lot of space for their numerous parameters (sample counts, binwidth), especially in multidimensional settings. Kernel density estimators also rely on suitable choices of basis functions, which yet again need to be periodic for the present task. As mentioned before, it is however possible to wrap any linear distribution around the unit circle by mapping subsequent intervals of (non-zero) length onto the interval $[0, 2pi)$ [96, 200]. Such distributions are then called *wrapped* distributions. Given a random variable with density function f on the real line, the density of the wrapped variable [34] is defined as

$$f_w(x) = \sum_{k=-\infty}^{+\infty} f(x \bmod 2\pi k) \tag{50}$$

with distribution

$$F_w(x) = \sum_{k=-\infty}^{+\infty} \left(F(x + 2\pi k) - F(2\pi k)\right) \ , \tag{51}$$

subject to

$$p(x) \geqslant 0 \tag{52}$$

$$p(x) = p(x + 2\pi) \tag{53}$$

$$\int_{x=0}^{2\pi} p(x) dx = 1 \ . \tag{54}$$

For example, the *Wrapped Normal* is obtained by wrapping the linear normal distribution as follows:

$$\mathcal{N}_w(x|\mu, \sigma) = \sum_{w=-\infty}^{+\infty} \mathcal{N}(x + 2\pi w|\mu, \sigma) \tag{55}$$

It can be shown [200] that mean and variance of a wrapped distribution are strictly related to their circular counterparts through

$$\mu_{circular} = \mu \mod 2\pi \quad \text{and} \quad \sigma^2 = -2\ln(1 - \nu) \ , \tag{56}$$

which in turn motivates the definition of the circular standard deviation as in equation (49). One may further note that equation (55) also closely approximates the density of the von Mises distribution, as illustrated in figure 14. According to Fisher [96], choosing between the von Mises and the Wrapped Normal is usually a matter of taste. Just like mixtures of von Mises distributions can be used to achieve multimodality, so can mixtures of Wrapped Normals. Recall that joint densities of periodic and linear random variables tend to become intractable. In this regard, wrapped linear distributions are preferable

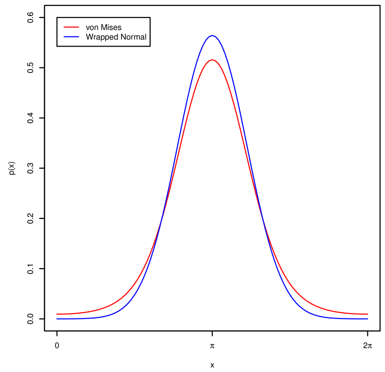


Figure 14.: The density of the von Mises distribution is closely approximated by the Wrapped Normal.

over periodic distributions [275, 21, 96]. Eventually, this suggests the use of mixtures of wrapped multivariate normals for the present dataset. In accordance with equation (55), the density of a *Wrapped Gaussian Mixture Model (W-GMM)* is defined as

$$p(\mathbf{x}|\boldsymbol{\theta}) = \sum_{k} \sum_{\boldsymbol{w} \in \mathbb{Z}^{D}} \mathcal{N}(\mathbf{x} + 2\pi\boldsymbol{w}|\boldsymbol{\mu}_{k}, \boldsymbol{\Sigma}_{k}) \tag{57}$$

More precisely, though, only the circular variables are modeled by Wrapped Normals whereas linear variables must remain as is, finally resulting in a so-called *Semi-Wrapped Gaussian Mixture Models (SW-GMMs)*, for which the above equation is thus slightly altered:

$$p(\mathbf{x}|\boldsymbol{\theta}) = \sum_{k} \sum_{\boldsymbol{w} \in \boldsymbol{W}} \mathcal{N}(\mathbf{x} + 2\pi\boldsymbol{w}|\boldsymbol{\mu}_{k}, \boldsymbol{\Sigma}_{k}) \tag{58}$$

Here, $\boldsymbol{W} \subseteq \mathbb{Z} \times \mathbb{Z} \times \ldots \times \mathbb{Z}$ denotes a set of D-dimensional displacement vectors, where the i-th element of $\boldsymbol{w}$ corresponds to the i-th random variable for all $\boldsymbol{w} \in \boldsymbol{W}$. It follows that those $\boldsymbol{w}_i$ that correspond to linear variables will remain 0 at all times.

2.3.2.3 *Approximating Wrapped Distributions*

In practice one can only approximate wrapped distributions because, depending on the number of wrapped variables, the actual number of displacements causes exponentational growth of the computational costs. For example, given a set of periodic variables V and a corresponding function $f : V \to \mathbb{N}_{+}$ which maps each variable to a selected number of tilings, the costs grow by a factor of $\prod_{v \in V} f(v)$. It follows that for the modeling of the dataset at hand, of which two out of three variables represent angular data, choosing as much as 5 displacements per periodic variable (i. e. $\boldsymbol{w} \in \{-2, -1, 0, 1, 2\}^{2} \times \{0\}$) would already scale the computational costs by a factor of 25. It should be noted that this factor can only serve as a lower bound due to additional system-architecture dependent bottlenecks like caches etc. As a consequence, it is most often suggested that a maximum of 3

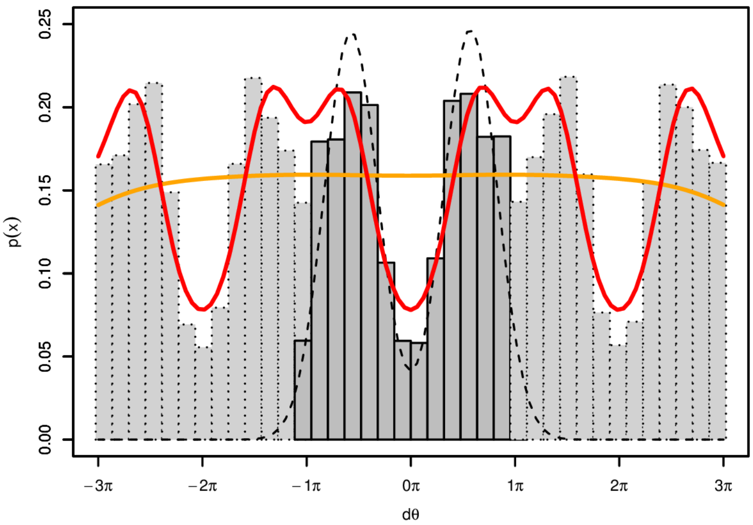


Figure 15.: The middle histogram shows the actual distribution of a subset of $\delta\theta$ over $[pi, pi)$, the left and right histograms show additional tilings. The density of a regular GMM is shown by the dashed line while the orange and red lines correspond to the densities of SW-GMMs with 2 and 4 components, effectively demonstrating the potential requirement for additional components for multimodal linear wrapped distributions.

tilings should be used per variable [10, 275, 21], while more than 6 tilings are generally considered intractable. On the other hand, the minimum required number of tilings depends on the actual distribution of the corresponding variable. The general consensus however is that an approximation using 3 tilings is legitimate for variables for which it holds that

$$3\sqrt{\sigma^2} \leqslant 2\pi \quad \Rightarrow \quad \sigma \leqslant \left(\frac{2}{3}\pi\right)^2 . \tag{59}$$

This is reasonable since, based on the *3 sigma rule*, even for larger variances of up to $(\frac{2}{3}\pi)^2$ at least 99.7% of the samples are located within ± 1 tilings around the mean. Depending on the actual distribution of the data, SW-GMMs typically require more modes than GMMs. This is a consequence of the fact that EM is based on maximization of the complete-data log likelihood, for which in case of SW-GMMs multiple tilings and hence more data are taken into account during the training phase. This is illustrated in figure 15.

2.3.2.4 *Learning Semi-Wrapped Gaussian Mixture Models*

The idea behind SW-GMMs and their algorithmic basics have been discussed in [275, 21], yet none of which gives an exhaustive treatment of EM. To the best of the author's knowledge, the following is the first complete derivation of EM for SW-GMMs.

Recall that for GMMs the latent variable $\boldsymbol{z}$ encodes the responsible mixture component in a 1-of-K coding. Now let $\boldsymbol{W} \in \mathbb{Z}^{P \times D}$ be a matrix whose p-th row represents the p-th tiling's displacement of the original D-dimensional samples. For this, let $\boldsymbol{w}$ be a latent variable with 1-of-P coding, corresponding to a single tiling. These variables are then subject to

$$z_k \in \{0,1\} \wedge \sum_{k=1}^{K} z_k = 1 \quad \mathit{respective} \quad w_p \in \{0,1\} \wedge \sum_{t=1}^{T} w_p = 1\ . \tag{60}$$

Define the joint density of the independent $\boldsymbol{z}$ and $\boldsymbol{w}$ as

$$\begin{aligned} p(\boldsymbol{z}_k = 1, \boldsymbol{w}_p = 1|\theta) &= \boldsymbol{\pi}_\mathbf{k} \qquad (61) \\ &= \prod_k \prod_t \boldsymbol{\pi}_\mathbf{k}^{z_k w_p}\ . \qquad (62) \end{aligned}$$

Likewise, the probability of a sample $\boldsymbol{x}$, given the respective component $\boldsymbol{z}_k$ and tiling $\boldsymbol{w}_p$, is defined as

$$p(\mathbf{x}|\boldsymbol{z}_k = 1, \boldsymbol{w}_p = 1, \theta) = \mathcal{N}(\boldsymbol{x} + 2\pi \mathbf{W_p} | \boldsymbol{\mu}_\mathbf{k}, \boldsymbol{\Sigma}_\mathbf{k})\ , \tag{63}$$

which due to the variables' encoding can be rewritten as

$$p(\mathbf{x}|\boldsymbol{z}, \boldsymbol{w}, \theta) = \prod_k \prod_t (\mathcal{N}(\boldsymbol{x} + 2\pi \mathbf{W_p} | \boldsymbol{\mu}_\mathbf{k}, \boldsymbol{\Sigma}_\mathbf{k}))^{z_k w_p}\ . \tag{64}$$

It follows that the joint density of $\boldsymbol{x}$, $\boldsymbol{z}$ and $\boldsymbol{w}$ is given by

$$p(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{w}|\theta) = \prod_k \prod_t (\boldsymbol{\pi}_\mathbf{k} \mathcal{N}(\boldsymbol{x} + 2\pi \mathbf{W_p} | \boldsymbol{\mu}_\mathbf{k}, \boldsymbol{\Sigma}_\mathbf{k}))^{z_k w_p}\ . \tag{65}$$

Verify that the shape of the original model was retained throughout the process:

$$\begin{aligned} p(\mathbf{x}|\boldsymbol{\theta}) &= \sum_z \sum_w p(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{w}) \qquad (66) \\ &= \sum_k \sum_t \boldsymbol{\pi}_\mathbf{k} \mathcal{N}(\boldsymbol{x} + 2\pi \mathbf{W_p} | \boldsymbol{\mu}_\mathbf{k}, \boldsymbol{\Sigma}_\mathbf{k}) \qquad (67) \end{aligned}$$

The above equations now allow to determine the *responsibilities* according to the posterior of $\boldsymbol{z}$ and $\boldsymbol{w}$, given $\boldsymbol{x}$ and $\boldsymbol{\theta}$

$$\begin{aligned} p(\boldsymbol{z}, \boldsymbol{w}|\mathbf{x}, \boldsymbol{\theta}) &= \frac{p(\mathbf{x}|\boldsymbol{z}, \boldsymbol{w}, \boldsymbol{\theta}) \cdot p(\boldsymbol{z}, \boldsymbol{w}, \boldsymbol{\theta})}{p(\mathbf{x}|\boldsymbol{\theta})} \qquad (68) \\ &= \frac{\prod_k \prod_t (\boldsymbol{\pi}_\mathbf{k} \mathcal{N}(\boldsymbol{x} + 2\pi \mathbf{W_p} | \boldsymbol{\mu}_\mathbf{k}, \boldsymbol{\Sigma}_\mathbf{k}))^{z_k w_p}}{\sum_{k'} \sum_{t'} \mathcal{N}(\boldsymbol{x} + 2\pi \mathbf{W_{t'}} | \boldsymbol{\mu}_\mathbf{k'}, \boldsymbol{\Sigma}_\mathbf{k'})}\ . \qquad (69) \end{aligned}$$

Alternatively, the responsibility of the k-th mixture component for explaining the p-th displacement of a given sample $\boldsymbol{x}_n$ is determined by

$$\gamma(\boldsymbol{z}_{nk}, \boldsymbol{w}_p) = p(\boldsymbol{z}_k = 1, \boldsymbol{w}_p = 1|\mathbf{x}, \theta) = \frac{\boldsymbol{\pi}_\mathbf{k} \mathcal{N}(\boldsymbol{x} + 2\pi \mathbf{W_p} | \boldsymbol{\mu}_\mathbf{k}, \boldsymbol{\Sigma}_\mathbf{k})}{\sum_{k'} \sum_{t'} \mathcal{N}(\boldsymbol{x} + 2\pi \mathbf{W_{t'}} | \boldsymbol{\mu}_\mathbf{k'}, \boldsymbol{\Sigma}_\mathbf{k'})}\ . \tag{70}$$

Then the expected complete-data log-likelihood for a set $\mathbf{X}$ of N i.i.d. samples is

$$\mathcal{Q}(\theta, \theta^{t-1}) = \mathbb{E}\left[\sum_{n} \ln p(x, z, w)\right] \tag{71}$$

$$= \sum_{n} \mathbb{E}\left[\ln \prod_{k} \prod_{t} (\pi_k \mathcal{N}(x + 2\pi \mathbf{W_p} | \mu_k, \Sigma_k))^{z_k w_p}\right] \tag{72}$$

$$= \sum_{n} \sum_{k} \sum_{t} \underbrace{\mathbb{E}\left[z_k w_p\right]}_{\gamma(z_{nk}, w_p)} \ln\{\pi_k \mathcal{N}(x + 2\pi \mathbf{W_p} | \mu_k, \Sigma_k)\} \ . \tag{73}$$

This leads to the update rules

$$\pi_k^{t+1} = \frac{N_k}{N} \tag{74}$$

$$\mu_k^{t+1} = \frac{1}{N_k} \sum_{n} \sum_{t} \gamma(z_{nk}, w_p)(x_n + 2\pi \mathbf{W_p}) \tag{75}$$

$$\Sigma_k^{t+1} = \frac{1}{N_k} \sum_{n} \sum_{t} \gamma(z_{nk}, w_p)\, \xi(x_n - \mu_k^{t+1} + 2\pi \mathbf{W_p})\ , \tag{76}$$

for which $N_k = \sum_n \sum_t \gamma(z_{nk}, w_p)$ and $\xi : v \mapsto vv^T$ maps a given vector to its outer product.

2.3.3 *Computing the models*

2.3.3.1 *Initialization*

Estimation of a suitable initial parameter set for either GMM or SW-GMM with K mixture components is done by applying the K-Means algorithm to the training data. K-Means is a hard-clustering algorithm that assigns each sample point to one of K cluster centers, and is in fact closely related to the EM algorithm for GMMs, the latter of which makes only soft assignments based on the posterior probabilities. This relation can be demonstrated by considering the limit $\epsilon \to 0$ of a GMM with covariances of the form $\epsilon \cdot \mathbf{I}$, where $\mathbf{I}$ denotes identity [34]. Once K has been carefully chosen, which will be further investigated in section 2.3.4, the K cluster centers found by K-Means can very well serve as an initial guess for the means of the K mixture components. Estimates for the covariances are then determined by computing the covariance matrices of each set of points assigned to the respective clusters. An initial estimate of the mixing coefficients is consequently given by the ratio of the size of each cluster to the total size of the training data.

It is important to note that K-Means is based on the Euclidean distance between points whereas GMMs – or, more specifically, normal distributions – are based on Mahalanobis distance and hence taking into account the variables' covariances. This means that K-Means will naturally perform poorly on datasets where the variances of the variables differ by one or more magnitudes. Moreover, this explains why K-Means is non-robust to outliers. For the present dataset, the measured interpersonal distances vary from about 194 mm to about 3025 mm with $\sigma \approx 429\ mm$ for $S^{\oplus}$ and $\sigma \approx 484\ mm$ for $S^{\ominus}$, as opposed to the 2π-periodic $\delta\theta$ and $\delta\varphi$ with much smaller standard deviations (which were given in table 8). Since it is vital that for K-Means all variables live on the same scale, these have to be transformed accordingly, e. g. by subtracting their mean and scaling to unit variance, known as *standardizing* or *feature scaling* [34].

2.3.3.2 *Whitening*

Prior to feature scaling, the present data were furthermore decorrelated for decreased redundancy [128] and reduction of noise. Decorrelation is also likely to improve the convergence characteristics due to the duality between the input space and the space of the error function, for which it is presumed that orthogonalization of the input space has some orthogonalizing effect on the error function as well, meaning that surface dents become more symmetric and hence their gradients easier to travel [236].
An effective way for orthogonalization and scaling is through Principal Component Analysis (PCA) [34, 185]. PCA is an orthonormal mapping of data onto their principal components and can be used for decorrelation and/or information reduction, the latter of which is achieved by selecting less components than the original number of dimensions. The principal components are typically chosen by determining a new set of orthonormal basis vectors which maximize variance. This can be achieved by maximizing the second central moment of $\mathbf{X}$ under the transformation $\mathbf{U}$

$$\mathbb{E}\left[(\mathbf{X}\mathbf{u}_i)^T(\mathbf{X}\mathbf{u}_i)\right] = \mathbb{E}\left[\mathbf{u}_i^T\mathbf{X}^T\mathbf{X}\mathbf{u}_i\right] = \mathbf{u}_i^T \underbrace{\mathbb{E}\left[\mathbf{X}^T\mathbf{X}\right]}_{=\mathbf{\Sigma}} \mathbf{u}_i\ , \tag{77}$$

where $\mathbf{u}_i$ denotes the i-th column of $\mathbf{U}$, and without loss of generality (w.l.o.g.) the data have zero mean. The orthogonality constraint on the basis vectors can be enforced by a Lagrange multiplier when the above equation is maximized through its derivative:

$$\frac{d}{d\mathbf{u}}\mathbf{u}^T\mathbf{\Sigma}\mathbf{u} + \lambda(\mathbf{u}^T\mathbf{u} - 1) \overset{!}{=} 0 \quad \Leftrightarrow \quad \mathbf{\Sigma}\mathbf{u} = \lambda\mathbf{u}\ . \tag{78}$$

Solving for the eigenpairs of $\mathbf{\Sigma}$ yields the eigenvalue decomposition of the covariance matrix. The eigenvalues correspond to the variances along the respective eigenvectors. Note that full PCA does not alter the sum of variances since $\mathbf{U}$ is orthonormal and the trace of a matrix is invariant under cyclic permutations:

$$\mathrm{tr}\left((\mathbf{X}\mathbf{U})^T(\mathbf{X}\mathbf{U})\right) = \mathrm{tr}\left(\mathbf{U}^T(\mathbf{X}^T\mathbf{X})\mathbf{U}\right) = \mathrm{tr}\left(\mathbf{U}^T\mathbf{\Sigma}\mathbf{U}\right) = \mathrm{tr}\left(\mathbf{U}\mathbf{U}^T\mathbf{\Sigma}\right) = \mathrm{tr}\left(\mathbf{I}\mathbf{\Sigma}\right) = \mathrm{tr}\left(\mathbf{\Sigma}\right)\ . \tag{79}$$

The scaling factor $1/(N-1)$ has been omitted as it does not contribute to the above equation.

Due to its numerical stability, computing the eigenpairs is preferably done via SVD of the *data* instead of eigenpair decomposition of the *covariance matrix.* SVD yields a decomposition of the form $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^\mathsf{T}$, where $\mathbf{U}$ and $\mathbf{V}$ are the left and right singular vectors of $\mathbf{X}$, and $\mathbf{D}$ is a diagonal matrix whose diagonal contains the singular values of $\mathbf{X}$. The left and right singular vectors are the eigenvectors of $\mathbf{X}\mathbf{X}^*$ respective $\mathbf{X}^*\mathbf{X}$, which means that for zero-mean data the right singular vectors are in fact identical to the eigenvectors of the covariance matrix. All the same, the singular values correspond to the square roots of the eigenvalues of the covariance matrix, because

$$(\mathbf{U}\mathbf{D}\mathbf{V}^\mathsf{T})(\mathbf{U}\mathbf{D}\mathbf{V}^\mathsf{T})^\mathsf{T} = (\mathbf{U}\mathbf{D}\mathbf{V}^\mathsf{T})(\mathbf{V}\mathbf{D}\mathbf{U}^\mathsf{T}) = \mathbf{U}\mathbf{D}^2\mathbf{U}^\mathsf{T}\,. \tag{80}$$

Due to the orthonormality of $\mathbf{U}$, this can be interpreted as a rotation into orthogonal space, followed by a per-axis scaling, and an inverse rotation back to input space. The fact that the diagonal of $\mathbf{D}$ equals the standard deviations of the (now orthogonal) variables makes it as well easy to scale the data to unit variance by replacing each element of $\mathbf{D}$ with its reciprocal, resulting in $\mathbf{D}^{-1}$. The final decorrelation and scaling transformation is thus given multiplication of the centered data with $\boldsymbol{W} := \mathbf{U}\mathbf{D}^{-1}$. Since Gaussians are parameterized only in terms of mean and covariance of the data, linear transformation of the data causes no harm because

$$\mathbb{E}[\mathbf{A}\mathbf{X} + \mathbf{b}] = \mathbb{E}[\mathbf{A}\mathbf{X}] \quad \mathit{and} \quad \mathrm{Cov}[\mathbf{A}\mathbf{X} + \mathbf{b}] = \mathbf{A}\mathbf{X}\mathbf{A}^\mathsf{T}\,. \tag{81}$$

The training of models such as mixtures of Gaussians GMMs on linearly transformed data $\hat{\mathbf{X}} = \mathbf{X}\boldsymbol{W}$ is therefore equivalent to training on the original data $\mathbf{X}$. The parameters of the resulting model can be transformed back to the original input space by computing the means of the mixture components

$$\mu_k = \hat{\mu_k}\mathbf{D}\mathbf{U}^\mathsf{T} + \overline{\mu}\,, \tag{82}$$

where $\overline{\mu}$ is the mean of the original input data $\mathbf{X}$, and likewise the covariance matrices as

$$\begin{aligned}\Sigma_k &= \frac{1}{N-1}\mathbf{X}^\mathsf{T}\mathbf{X} \\ &= \frac{1}{N-1}(\hat{\mathbf{X}}\boldsymbol{W}^{-1})^\mathsf{T}(\hat{\mathbf{X}}\boldsymbol{W}^{-1}) \\ &= \boldsymbol{W}^{-\mathsf{T}}\underbrace{\frac{1}{N-1}\hat{\mathbf{X}}^\mathsf{T}\hat{\mathbf{X}}}_{\hat{\Sigma}}\boldsymbol{W}^{-1} \\ &= (\mathbf{D}\mathbf{U}^\mathsf{T})^\mathsf{T}\hat{\Sigma}\mathbf{D}\mathbf{U}^\mathsf{T}\,.\end{aligned} \tag{83}$$

2.3.3.3 *Computing in log-space*

Accumulation and/or multiplication of multiple very small probabilities may quickly exceed the numerical range of floating point architectures. Cancellation can e. g. be avoided by scaling [255], where probabilities or other inferred entities are scaled in a way such that the scaling coefficients cancel out only once a computation is finished. Another common approach is to perform all calculations in log-space, the downside of which is a notable increase in computational overhead. For operations in log-space, [199] suggests the use of an *extended logarithm*, where

$$\mathsf{eln}(x) = \begin{cases} \log(x) & \text{if } x > 0 \\ \mathit{LZERO} & \text{if } x = 0 \end{cases} \quad \text{and} \quad \mathsf{eexp}(x) = \begin{cases} \exp(x) & \text{if } x > 0 \\ 0 & \text{if } x = \mathit{LZERO} \end{cases} \tag{84}$$

and *LZERO* is defined as either `NaN` or $-\infty$, depending on architecture. It is furthermore vital to define an appropriate *sum operator* to compute $\mathsf{eln}(x+y)$, given $\mathsf{eln}(x)$ and $\mathsf{eln}(y)$. This operator should be defined in a way such that exponentation is avoided or else only used in a numerically stable manner. The sum operator $\oplus_L$ is accordingly defined as

$$\begin{aligned} \mathsf{eln}(x) \oplus_L \mathsf{eln}(y) &= \mathsf{eln}(x+y) \\ &= \mathsf{eln}(x) + \mathsf{eln}(x+y) - \mathsf{eln}(x) \\ &= \mathsf{eln}(x) + \mathsf{eln}\left(\frac{x+y}{x}\right) \\ &= \mathsf{eln}(x) + \mathsf{eln}\left(1 + \frac{y}{x}\right) \\ &= \mathsf{eln}(x) + \mathsf{eln}(1 + \mathsf{eexp}(\mathsf{eln}(y/x))) \\ &= \mathsf{eln}(x) + \mathsf{eln}(1 + \mathsf{eexp}(\mathsf{eln}(y) - \mathsf{eln}(x))) \ . \end{aligned} \tag{85}$$

For further increase in numerical stability, values are kept small by swapping $\mathsf{eln}(x)$ and $\mathsf{eln}(y)$ whenever $\mathsf{eln}(x) > \mathsf{eln}(y)$. Similar to $\oplus_L$, a product operator for $\mathsf{eln}(x \cdot y)$, given $\mathsf{eln}(x)$ and $\mathsf{eln}(y)$, is defined as

$$\mathsf{eln}(x) \odot_L \mathsf{eln}(y) = \mathsf{eln}(x \cdot y) = \begin{cases} \mathsf{eln}(x) \oplus_L \mathsf{eln}(y) & \text{if } x > 0 \wedge y > 0 \\ \mathit{LZERO} & \text{if } x = 0 \vee y = 0 \end{cases} . \tag{86}$$

APPLICATION TO EM The defined operations can be used throughout EM for SW-GMMs. It is only consequent (and faster) to compute the responsibilities of Gaussian components based on log-space probabilities as well. For this, consider the logarithm of the probability density function for a multivariate Gaussian

$$\log p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \log\left(\frac{1}{\sqrt{(2\pi)^D |\boldsymbol{\Sigma}|}} \cdot \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})\right)\right) , \tag{87}$$

which can be rewritten as

$$\log p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{1}{2}\left(C + \log|\boldsymbol{\Sigma}| + (\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})\right) \tag{88}$$

where $C = D \cdot \log(2\pi)$. Both the logarithm of the determinant and the Mahalanobis distance can be very efficiently computed by means of QR decomposition. So most if not all computations can be performed in log-space while the computational demand is kept at bay. This is especially important due to the large amount of samples in the present dataset, along with the previous considerations about the necessary number of tilings and the respective exponential growth of the computational costs (discussed in sections 2.3.2.3 and 2.3.2.4).

2.3.3.4 *Avoiding singularities*

EM maximizes the log-likelihood of the complete data. A Gaussian with its center on a single sample and with zero variance maximizes the probability of that sample being drawn from the corresponding Gaussian, and hence maximizes the overall log-likelihood. Implementations of the EM algorithm must therefore take care to avoid these singularities for modeling and numerical reasons. In general, this problem can be completely avoided by using Variatonal Mixture of Gaussians [34], based on a fully Bayesian model with prior distributions over the whole set of parameters, including the number of components. It is precisely because of these prior distributions that singularities do not occur. In addition to that, the maximum likelihood number of components can be inferred probabilistically. On the other hand, the downside of this approach is yet again increased complexity.
A much simpler approach is to keep track of the determinants of the covariance matrices, for which values close or equal to zero indicate that a particular Gaussian is about to collapse. Whenever that happens, the routine could reinitialize that Gaussian's parameter set with random values or restart the whole learning process with a different initial setup. Moreover, the superimposition of noise onto the model parameters at each iteration may both avoid singularities and help EM to pass through shallow local optima of the target function.

2.3.4 *Model selection*

Model selection aims at finding the best model for a given dataset. For non-probabilistic models, such as K-Means, the best model is usually found by minimization of the reconstruction error. For probabilistic models, however, the best model may be found by cross-validating a set of models and choosing the one with the best fit for the data. A more efficient approach [218] is based maximizing the posterior of a model m, given the data $\mathcal{D}$:

$$p(m|\mathcal{D}) = \frac{p(\mathcal{D}|m) \cdot p(m)}{\sum_m p(m, \mathcal{D})} \tag{89}$$

Assuming equal prior $p(m)$ for all models, this is then equivalent to maximizing

$$p(\mathcal{D}|m) = \int p(\mathcal{D}|\theta)p(\theta|m)d\theta \; . \tag{90}$$

In order to avoid the (potentially complex) evaluation of the integral in equation (90), approximations like e. g. the Bayesian Information Criterion (BIC) are commonly used instead. BIC assesses the maximum likelihood estimate $\hat{\theta}$ in relation to the model's degrees of freedom. Given a set of N i.i.d. observations and a set $\hat{\theta}$ of model parameters with K degrees of freedom, BIC penalizes overly complex models:

$$\text{BIC} = \log p(D|\hat{\theta}) - \frac{K}{2} \cdot \log N \; . \tag{91}$$

Other than BIC, the Akaike Information Criterion (AIC) is not based on the marginal likelihood, but rather inferred from a frequentist perspective [218]. The definition of the AIC is quite similar to that of the BIC, but does not take into account the number of samples. Its penalty term is generally less when compared to the BIC. As such, the AIC suffers from a tendency to prefer more complex models. Therefore, the Akaike Information Criterion corrected (AICc) imposes an additional penalty for extra parameters in relation to the finite number of samples:

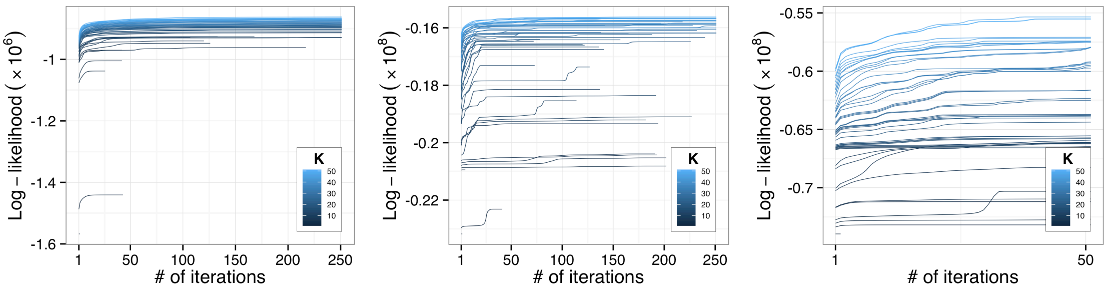
$$\begin{aligned} \mathrm{AICc} &= \mathrm{AIC} + \frac{2K(K+1)}{N-K-1} \\ &= \log p(D|\hat{\theta}) - K + \frac{2K(K+1)}{N-K-1} \end{aligned} \tag{92}$$

2.3.4.1 *Number of components*

An optimal number of mixture components must be carefully selected for both GMMs and SW-GMMs. For the latter, the number of additional tilings due to the periodic variables have to be taken into account. As a rule of thumb, the number of mixture components needs to increase along with the number of tilings. This is a consequence of maximum likelihood estimation, as it naturally attempts to find an optimal parameter set for explaining the *whole* dataset, which consequently involves the additional samples from the displaced periodic tilings. If the number of mixture components were not increased, part of the components would tend to exhibit significantly greater variance so as to be able to explain those samples that lie close to the limits of the domain. This is usually not an issue around the center of the distribution, but components with high variance gain more importance with increasing distance from the center, thus likely causing more misclassifications within these areas. Since SW-GMMs are only approximations of the real distribution (see section 2.3.2.3), further attention must be paid when selecting the number of components, as components with overly high variance can render wrapped distributions illegitimate, violating the constraint $\int_0^{2\pi} p(x)dx = 1$ for periodic distributions.

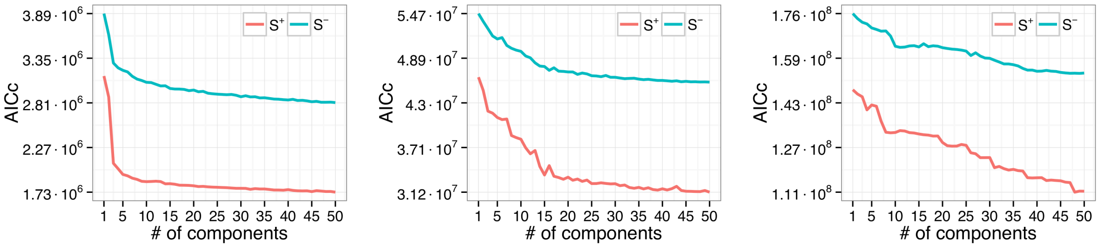
2.3.5 *Evaluation*

Following the discussions in 2.3.4 and 2.3.2.3, various configurations of models were computed and their characteristics were compared. The parameter settings varied among the number of mixture components as well as the number of tilings for SW-GMMs. Figure 16 illustrates the convergence characteristics for GMMs and SW-GMMs on the $S^{\oplus}$ dataset. For this purpose, only those models with either one or two wraps for *both* periodic variables have been selected as representatives from the various possibilities. As expected, EM converges relatively fast and straight-forward for regular GMMs, while there is much less progress for the more complex SW-GMMs at each iteration, since estimating the latter involves evaluating 9 to 25 times more samples from the obligatory tilings. Even more so,



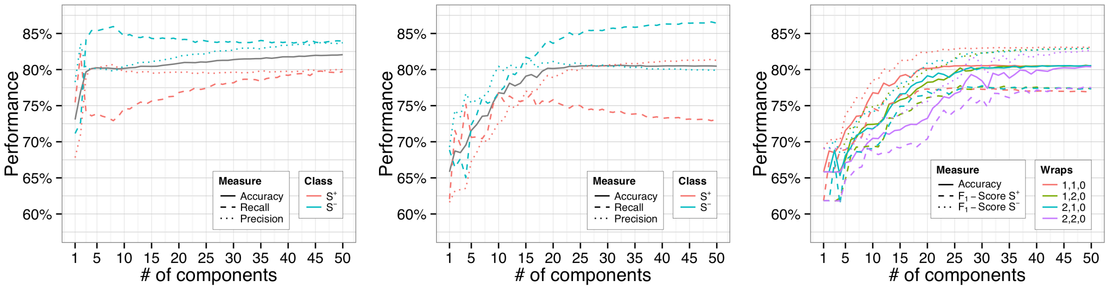
(a) (b) (c)

Figure 16.: Convergence characteristics of GMMs (a) and SW-GMMs with 1 (b) respective 2 (c) wraps per periodic variable on the $S^{\oplus}$ dataset.



(a) (b) (c)

Figure 17.: Information criteria of GMMs (a) and SW-GMMs with 1 (b) respective 2 (c) wraps per periodic variable.



(a) (b) (c)

Figure 18.: Performance characteristics of GMMs (a) and SW-GMMs (b) with 1 wrap per periodic variable. Comparison of SW-GMMs with varying number of wraps (c).

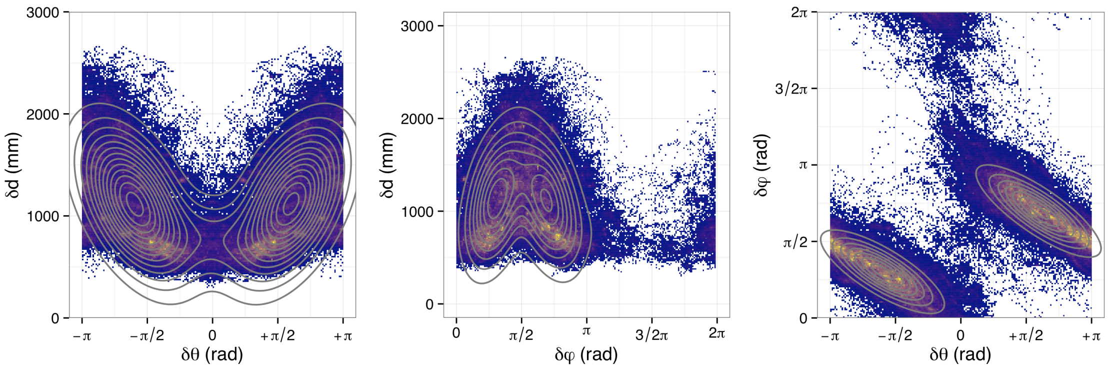
most GMMs converge towards a supposed global optimum, whereas SW-GMMs exhibit a tendency of ending up in local optima. This effect becomes less with an increase in the number of components, which is in agreement with the prior discussion.

AICc further corroborates this argument (figure 17). The quality of the models increases with the number of components, although a considerable saturation effect can be seen for GMMs with as few as five to ten components, after which no change of notable magnitude is to be expected. The saturation is not so pronounced for SW-GMMs, especially when more than ± 1 tilings are involved. In the depicted range of up to fifty Gaussians the penalty term of AICc does not yet reveal overfitting from a information-theoretical perspective. In fact this is even not the case when computing mixture models with up to 150 components, yet models of that magnitude exhibit considerable spikes in the surfaces of their probability density functions, which clearly indicates overfitting and would contradict the premise of finding a preferrably simple, generalizable and interpretable model.

Looking at the log-likelihood and AICc from yet another point of view leads to the notion that GMMs explain the data much better than their periodic siblings. This naturally raises the question for what may be the cause, especially after the expected benefits of (semi-)wrapped distributions for the present setting involving linear and periodic variables. Arguably, this can be explained for three reasons: First, when comparing GMMs and SW-GMMs with the same number of components, the latter naturally need to be comprised of Gaussians with higher variance in order to compensate for the additional samples in the periodic tilings, and hence each of the samples becomes less probable. Second, apart from accumulated probabilities, log-likelihood is also a function of the number of samples, so that the additional data from the tilings again have strong influence on the overall likelihood. Third, and this is the most interesting aspect, it turns out that the *actual* distribution of the samples (refer to figure 7) is such that it can be approximated quite well and with very reasonable accuracy by a regular GMM. For illustration purposes, consider a trivial model comprised of only 3 components. Figure 19 shows that, in a qualitative sense, the natural shape of the Gaussians overcomes the shortcoming of GMMs to see data beyond periodic borders, if only to a certain extent. This hypothesis is sustained by reviewing the constraint $\int_0^{2\pi} p(x)dx = 1$ for periodic distributions. Evaluation of the respective integrals for the marginal and joint densities of a slightly more complex GMM, i. e. one with 10 components learned from the actual dataset, reveals that the integrated volume is in fact close to 1 (table 9). Nevertheless, there are non-negligible periodic characteristics in the data which are unlikely to be captured by GMMs with fewer modes. For example, refer to the samples in the area of $[-\frac{\pi}{2}, +\frac{\pi}{4}] \times [\frac{3}{2}\pi, 2\pi)$ in figure 19c. Apparently, this issue may be

	$p(\delta\theta, \delta\varphi)$	$p(\delta\theta)$	$p(\delta\varphi)$
$S^{\oplus}$	0.97	0.98	0.99
$S^{\ominus}$	0.95	0.97	0.98

Table 9.: Numerical quadrature over 2π-periodic intervals of the joint or marginal probability density functions of $\delta\theta$ and $\delta\varphi$, given a GMM with 10 components.



(a)

(b)

(c)

Figure 19.: Joint distributions of $\delta\theta$, $\delta\varphi$ and δd, superimposed with a contour plot of the probabilty density of a 3-Gaussians mixture model.

overcome with a slight increase in the number of components. As soon as too many Gaussians are chosen, though, models will at least tend to overfit the marginal $\delta\theta$. Comparison of the classification performance of the various configurations further suggests the selection of GMMs over SW-GMMs. Figure 18 shows that GMMs perform well, both in terms of overall classification accuracy, as well as precision and recall for both $S^{\oplus}$ and $S^{\ominus}$. Accuracy mostly lies above 80%, recall shows that 75% to 80% of social interactions are recognized as such, and only about 20% of the data were false positives for $S^{\oplus}$. Similar results can be seen for $S^{\ominus}$, although recall is slightly better for observations from $S^{\ominus}$. Comparable performance already starts at configurations with as few as 5 components. To achieve similar performance, SW-GMMs need way more components (as is expected). The recall of $S^{\oplus}$ and $S^{\ominus}$ is closer to the classifier's accuracy for SW-GMMs, but once overall accuracy reaches a satisfiable level of about 80%, the recall of $S^{\oplus}$ is getting worse. In order to get a notion of how these performance characteristics develop for varying numbers of tilings, figure 18c illustrates accuracies and F_1-scores for corresponding SW-GMMs. F_1-scores have been chosen instead of precision and recall to avoid further obfuscation of the graph.

So far, all models perform almost equally well, only some need to be more complex than others to achieve the same quality in performance. There is no apparent advantage of preferring SW-GMMs over GMMs. While SW-GMMs are certainly more correct in a theoretical sense, the question is whether their inherent additional complexity is justifiable or valuable enough for the present application domain. At the bottom line, the goal was finding a preferably general and thus not overly complex generative model, which was supposed to be explainable, updateable, and adaptable. Of course the model should perform well in classification tasks. With respect to their performance characteristics, and especially in regard of the fact that the present dataset represents social interaction of groups of various cardinalities and formations, both GMMs and SW-GMMs can certainly be considered as generalizing models. Depending on the choice of parameters, they are very well explainable, and they are adaptable in the aforementioned sense by their very nature. The proposed model therefore is one consisting of two GMMs, one for $S^{\oplus}$ and one for $S^{\ominus}$ with ten components each. To a certain extent, the specific choice of ten components is arguably somewhat arbitrary, but it yields a good compromise between having a universally applicable model and the ability to recognize rather specific effects from interaction geometry in human behaviour. Once more experimental data will be collected, the distributions of the samples for $S^{\oplus}$ and $S^{\ominus}$ will eventually converge towards their real distribution, which will presumably emphasize the more general aspects of proxemics, yet attenuate the remainder.

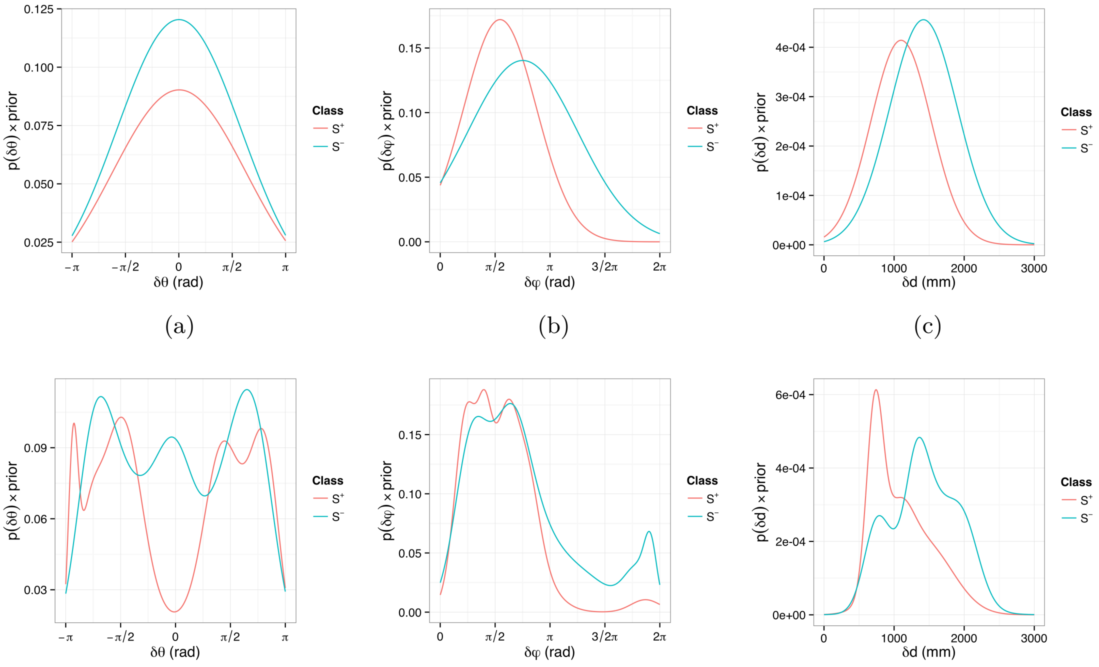
2.3.5.1 *Model performance versus other classifiers*

A comparison of the selected model's performance against other standard classifiers from [134] supports the choice of GMMs. According to table 10, none of the tested classifiers performs better in a way that would e. g. justify trading interpretability and simplicity for increased accuracy. Most notably, GMMs actually perform best next to SVMs. Interestingly enough, the simple Naïve Bayes exhibits about 72% in classification accuracy and, except for precision for $S^{\oplus}$, also shows reasonable quality for all other performance measures. This is noteworthy because Naïve Bayes not only assumes independence for each of the variables, but also that each of them corresponds to a single Gaussian. For the present two-class classification problem, this means modeling $\delta\theta$, $\delta\varphi$ and δd in terms of two univariate Gaussians each, one for $S^{\oplus}$ and for $S^{\ominus}$, effectively reducing information to as few as 6 model parameters for the Gaussians and 2 for the class priors. By doing so, Naïve Bayes simply bisects the variables' domains such that it e. g. considers observations $5° \leqslant \delta\varphi \leqslant 135°$ or $\delta d \leqslant 1185\,mm$ as $S^{\oplus}$ (see figure 20), and it is obviously biased towards $S^{\ominus}$. The drawbacks of erroneously assuming independence also show in the results for the more sophisticated forms of Naïve Bayes [42], as e. g. in the case of kernel density estimation (K) or supervised discretization (SD). Although the latter resemble the actual distributions in much more detail, a gap of almost 10% of performance still remains in comparison to GMMs or SVMs. Decision Trees, on the other hand, do not suffer from the independency assumption. The classification performance of the correspondingly chosen classifier is marginally better than that of Naïve Bayes. This is in part due to the selection of the parameters for pruning the estimated tree. These have been chosen such that there are at least 25,000 samples per leaf in order to avoid overfitting, resulting in a rudimentary explanation of the data (see appendix 62). Failure to do so also results in an overly deep decision tree, potentially contradicting the premise of generalizability.

All in all, GMMs perform very well per se, but particularly also in comparison to other standard classifiers. In the chosen configuration they are much less susceptible to overfitting than other models, yet still show way-above-average recall and precision for $S^{\oplus}$ and $S^{\ominus}$. According to the prior discussion, the choice of the number of components is a compromise between generalizability and handling presumed specific effects in the data. This is corroborated by the fact that the model allows for distinction of $S^{\oplus}$ and $S^{\ominus}$ even in settings of varying group sizes and corresponding sample variance. Once more experiments will be performed, which should preferably be conducted in the socio-psychological

		$S^{\oplus}$			$S^{\ominus}$		
Classifier	Acc.	Prec.	Rec.	F_1-Score	Prec.	Rec.	F_1-Score
Naïve Bayes	72.1%	66.8%	74.5%	70.5%	77.4%	70.3%	73.6%
Naïve Bayes (K)	73.6%	74.4%	62.5%	67.9%	73.2%	82.6%	77.7%
Naïve Bayes (SD)	73.5%	74.2%	62.3%	67.7%	73.1%	82.6%	77.6%
Decision Tree	74.7%	72.3%	70.1%	71.2%	76.5%	78.4%	77.5%
Logistic Regression	71.6%	69.6%	64.8%	67.1%	73.1%	77.2%	75.1%
Neural Network (1HL)	71.2%	67.1%	69.4%	68.3%	74.7%	72.6%	73.6%
SVM	81.4%	80.3%	77.3%	78.8%	82.3%	84.7%	83.5%
GMM	80.3%	80.1%	74.3%	77.1%	80.5%	85.1%	82.7%
SW-GMM	75.9%	71.3%	76.8%	74.0%	80.0%	75.1%	77.5%

Table 10.: Classifier performance on 10-fold stratified cross-validation. K, SD and 1HL denote the use of kernels, discretized values, and a single hidden layer, respectively. GMMs and SW-GMMs with 10 components each, and ±1 tilings per periodic variable.



(a)

(b)

(c)

(d)

(e)

(f)

Figure 20.: Posteriors of δθ, δφ and δd from Naïve Bayes [134] as opposed to the selected GMM.

fields of research, the specific choice of the number of components will have to be adapted according to potential changes in the distributions of either one or both of $S^{\oplus}$ and $S^{\ominus}$, provided that conducting further experiments and gathering more data would exhibit new or reshape existing clusters in the present data. Depending on the distribution of the samples in a further growing dataset, the constraints towards periodic distributions will likely not hold anymore for GMMs, and SW-GMMs will have to be reconsidered.

2.3.5.2 *How well does the model represent the data?*

The models for $S^{\oplus}$ and $S^{\ominus}$ show very good convergence towards the actual sample distribution of the data. Figure 21 illustrates the joint densities of $\delta\theta$, $\delta\varphi$ and δd, where all left-hand plots correspond to $S^{\oplus}$ and right-hand plots to $S^{\ominus}$. From figure 21a it follows that joint observations of $\delta\theta$ at slightly less than $90°$ and δd at $\sim 750\ mm$ constitute a strong indicator for the presence of social interaction. In accordance with the marginal of $\delta\theta$ (refer to figure 8a) a minor sink can be seen next to this area from which the probability then again increases along with distance and angles up to $180°$. This is yet another hint which supports the hypothesis that full frontal configurations are usually avoided at close distance, while they are increasingly common in FFS of larger groups. The emphasis of the former two Gaussians is also due to the marginal of δd (refer to figure 8e). As far as $S^{\ominus}$ is concerned, from figure 21b one can see that the joint distribution of $\delta\theta$ and δd is much more attenuated, particularly so at δd below 1000 mm. Similar to the distribution for $S^{\oplus}$, probability increases with distance and angle. As seen from $S^{\ominus}$, another notable discrepancy between the two distributions is given by means of a single Gaussian at $0°$ and about 1000 mm for $S^{\ominus}$, representing a configuration where one person faces the back of another person at relatively short distance. The lack of a corresponding Gaussian for $S^{\oplus}$ expresses its importance for characterizing the absence of social interaction.

Likewise, the joint density of $\delta\varphi$ and δd is characteristic and expressive for $S^{\oplus}$, and differs significantly from that of $S^{\ominus}$ (figures 21c and 21d). Two Gaussians at about $\pm 45°$ and 750 mm clearly indicate typical relative positions in mutual interaction, followed by two less expressive yet still important Gaussians at $\pm 10°$ and 1000 mm. Yet another Gaussian represents formations where one persons stands in front of another at δd above 1000 mm. The variance of all of these Gaussians increases the farther they are from 0 distance. Between angles of $\frac{3/2}{\pi}$ and 2π a slight increase can be seen below 1000 mm, accounting for the few observations at the rear, when one group member for example shortly turned to face another member of the group. When compared to the distribution for $S^{\ominus}$, the latter Gaussian is more or less insignificant, especially when taking into account the class priors, but it sustains the viability of GMMs for the distribution of the periodic variable at hand. Contrary to $S^{\oplus}$, the model for $S^{\ominus}$ has its peaks mostly beyond 1000 mm in the frontal hemisphere. A number of Gaussians account for the absence of social interaction in the rear, more precisely $\delta\varphi \in [\pi, 2\pi)$, particularly at distances of more than 750 mm. Two pairs of Gaussians at about $\{frac54\pi\, \frac{7.5}{4}\pi\} \times \{750, 1000\}$ emphasize typical formations for $S^{\ominus}$ which appear in $S^{\oplus}$ with much less concentration.

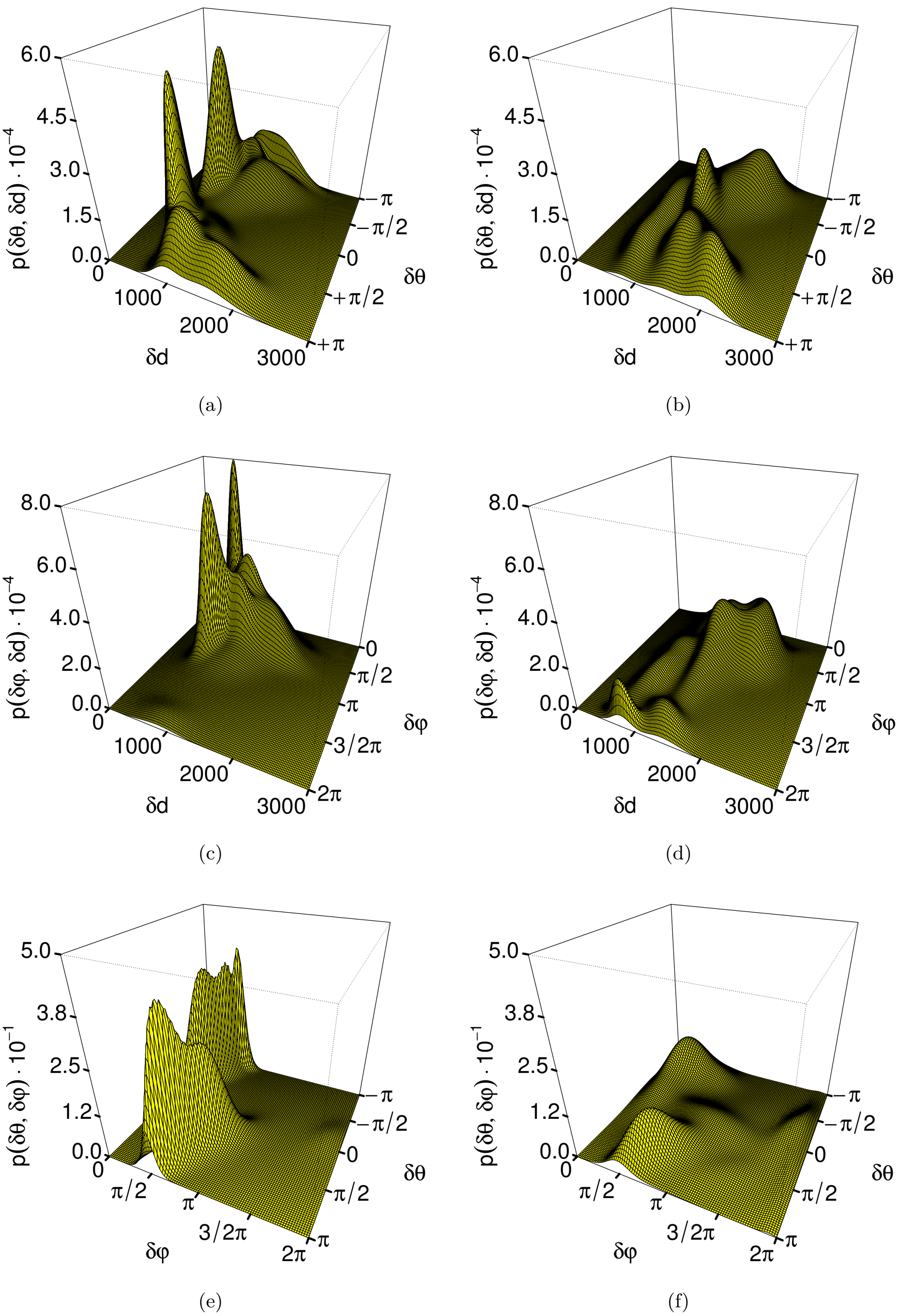


Figure 21.: Joint densities of the selected 10-components mixture models for $S^{\oplus}$ (a,c,e) and $S^{\ominus}$ (b,d,f).

	Predicted				
Actual	$S^{\oplus}$	$S^{\ominus}$	Precision	Recall	F_1-Score
$S^{\oplus}$	274737	93497	79.6%	74.6%	77.0%
$S^{\ominus}$	70253	387065	80.5%	84.6%	82.5%

(a) GMM

	Predicted				
Actual	$S^{\oplus}$	$S^{\ominus}$	Precision	Recall	F_1-Score
$S^{\oplus}$	282775	85459	72.7%	76.8%	74.7%
$S^{\ominus}$	106141	351177	80.4%	76.8%	78.6%

(b) SW-GMM

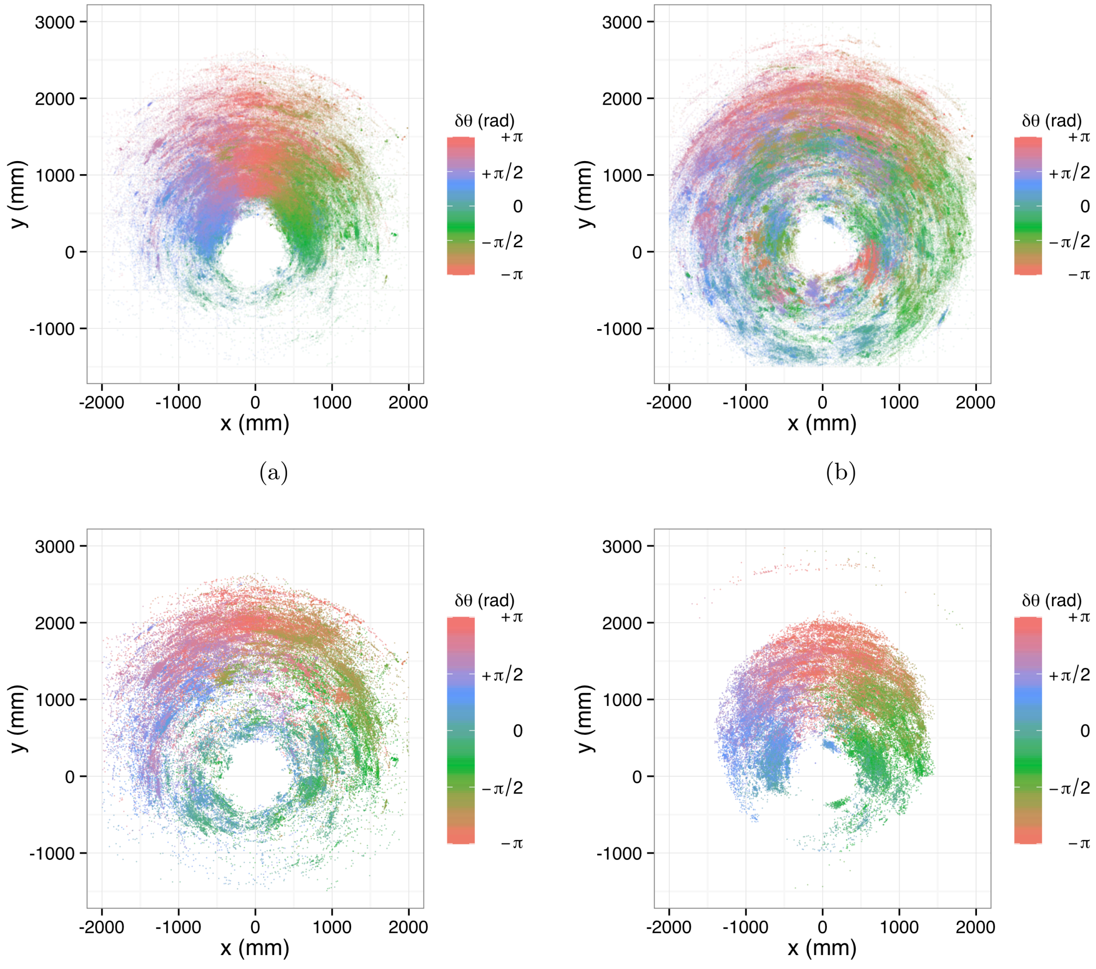
Table 11.: Confusion matrices after 10-fold stratified cross-validation of GMM- and SW-GMM-based classifiers.

Lastly, figures 21e and 21f reveal very distinct joint densities of $\delta\theta$ and $\delta\varphi$ for $S^{\oplus}$ and $S^{\ominus}$. The correlation of $\delta\theta$ and $\delta\varphi$, and their expressiveness for $S^{\oplus}$, is obvious. Once more, though, it is interesting that $\delta\theta$ and $\delta\varphi$ are clearly not independent from each other for $S^{\ominus}$. The increased number of observations at $(\frac{\pi}{2}, \frac{\pi}{2})$ and $(\frac{7}{4}\pi, \frac{2.5}{4}\pi)$, as well as the lack of observations in the rear, support the hypothesis that, in the absence of social interaction, configurations are not random at all. Further experiments would therefore lead to more emphasis on these effects, and consequently make the distinction of $S^{\oplus}$ and $S^{\ominus}$ more clear.

2.3.5.3 *Analysis of the classification results*

Although the presented classifier overall yields reasonable performance, its decision boundaries should be further explored. This may also give further insight into why recall and precision are slightly better for $S^{\ominus}$ than for $S^{\oplus}$, as can also be seen in detail from the confusion matrices in table 11. To a certain extent, this is possibly caused by the relative scale of the class priors. The number of observations of $S^{\ominus}$ is higher than that of $S^{\oplus}$, which is why the classifier will generally decide in favor of $S^{\ominus}$ in areas where there is no significant evidence for a particular class in the model. This is e. g. most likely to happen at greater distances or particularly so for observations in the rear hemisphere.

Figure 22 gives more insight into the classifier's decision boundaries by comparing an orthographic projection of the samples of each class of the whole dataset (figures 22a and 22b) versus those samples that were erroneously classified as false negatives or false positives (figures 22c and 22d). The data are projected from polar onto cartesian coordinates and their corresponding values of $\delta\theta$ are encoded through a color gradient. They hence represent an orthographic view of the observations as seen by a virtual single person located at the origin of the cartesian coordinate system.



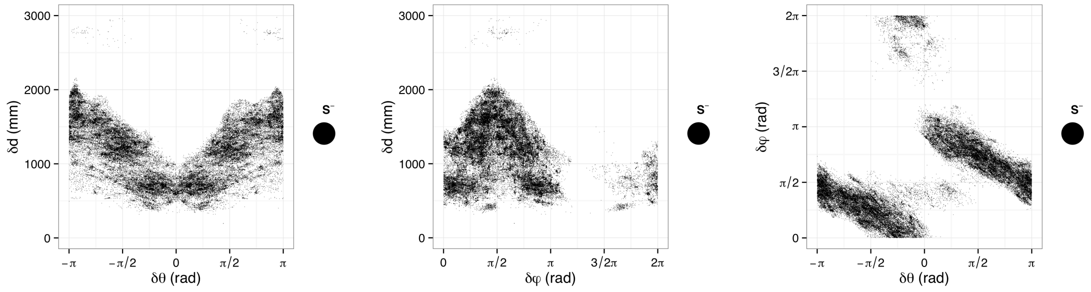
(a)

(b)

(c)

(d)

Figure 22.: Orthographic projection of the observations of $S^{\oplus}$ (a) and $S^{\ominus}$ (b) from the whole dataset vs. the misclassified observations from $S^{\oplus}$ (c) and $S^{\ominus}$ (d).



(a)

(b)

(c)

Figure 23.: Joint distributions of δθ, δφ and δd for false positives.

It is interesting to see how the $S^{\oplus}$ data form clusters in the shape of sectors of approximately equivalent shoulder orientations. Six such sectors can easily be spotted, namely those colored in blue, magenta, orange, brown, bright and dark green, three sectors of which seem dominant, i. e. blue, orange, and dark green. One can also see that variance increases with distance, especially beyond 1000 mm. As expected, there are only very few observations in the rear hemisphere. Still, it should be noted that these observations exhibit values of $\delta\theta$ representing shallow differences in orientation of the shoulder line, i. e. $|\delta\theta| \leqslant \frac{\pi}{2}$. Once more, this provides evidence towards the correctness of assuming that those observations of $S^{\oplus}$ in the rear hemisphere are mostly caused by one member of a group shortly turning towards another. Lastly, there is a noticable convex gap between $\sim 45°$ and $\sim 135°$, reaching out as far as 750 mm in front, where one might assume data due to the otherwise circular shape of the white area around the origin. This supports [308] who states that people are more likely to allow others to approach from the rear or the sides than from the front, and thus provides a refined view of the shape of the intimate and personal zones as suggested by Hall [133]. While the fact whether we accept “intruders” into our intimate or personal space is certainly also a function of social context, e. g. when acting in a rather crowded environment, comparison of figures 22a and 22b augments this view with a notion that such acceptance depends on mutual orientation of the bodies as well. For example, people generally demonstrate obvious annoyance in full-frontal configurations at close distances even in crowded scenarios, but tend to less strong reactions when facing the back of another person, for example when standing in a crowd and looking into the same direction, or when other persons may be passing by temporarily.
In contrast to $S^{\oplus}$, the data for $S^{\ominus}$ are much more irregular, which comes to no surprise. One might argue that, from a bird’s eye perspective, a rough shape of certain clusters in $\delta\theta$ is still recognizable. This is an effect that cannot be seen from the marginal of $\delta\theta$ (figure 8b) or any of the histograms of the joint distributions involving $\delta\theta$ (figures 7b and 7f) for $S^{\ominus}$, and is probably caused by the constrained movement area during the experiment. On the other hand, smaller size clusters of similar values of $\delta\theta$ build up everywhere throughout the domain, so that ultimately this matter has no recognizable effect on the generality of the model.

A qualitative view of the distribution of the misclassified samples, as shown in figures 22c and 22d, leads to the intuition that each of these distributions were just the opposite from those for the whole dataset, i. e. the distribution of the false negatives (figure 22c) looks like the overall distribution of $S^{\ominus}$ (figure 22b), and that of the false positives (figure 22d) like the overall distribution of $S^{\oplus}$ (figure 22a). This is a good result because it reveals a notion of *inversion* between the models for present and absent social interaction. On the downside, it shows a principle limitation of this approach towards modeling static social situations, namely that it can only provide a context-free interpretation of interaction geometry, not taking into account e. g. a time series of $\delta\theta$, $\delta\varphi$ and δd, or other potential sources of evidence for or against interaction. However, the consequences are much less for $S^{\oplus}$ than for $S^{\ominus}$. One is ultimately interested in detecting the presence of social situations, but not their absence, where the probability of encountering $S^{\oplus}$ is not strictly one minus

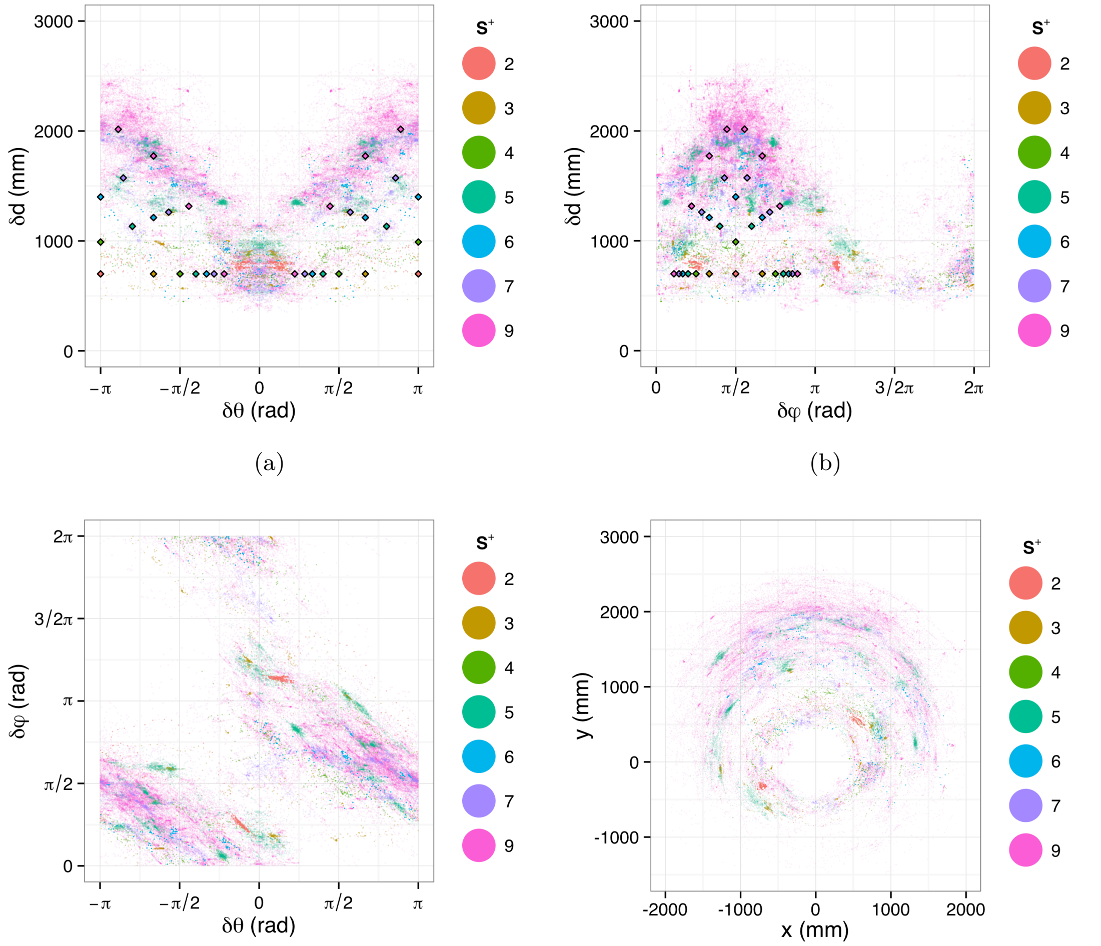
the probability of encountering $S^{\ominus}$. From figure 22d one can see that the decision boundary for $S^{\ominus}$ is semi-elliptical in terms of cartesian coordinates. Along the ordinate, observations above +2000 mm are generally classified correctly, and so are most below -500 mm. Along the abscissa, observations beyond $\pm 1300 \text{ mm}$ are correctly classified as well. The whole dataset contains more observations in the rear for $S^{\ominus}$ than it does for $S^{\oplus}$, and so the classifier is clearly biased towards $S^{\ominus}$ in that area. This is all the same done with respect to $\delta\theta$. The noticeable "hole" in the lower left quadrant lacks those observations from $S^{\ominus}$ where other persons stood rather close and had their shoulder lines oriented such they were more or less facing the same direction, which is yet another example of members of a group temporarily turning into the direction of other interactants. The same effect applies to the lower right quadrant, but is much less expressive in that area.
Other than that, a few false positives can be seen between 2000 mm and 3000 mm. This is a consequence of deciding in favor of present social interaction in case of equal posteriors of the models for $S^{\oplus}$ and $S^{\ominus}$, which are both zero for the aforementioned observations due to numerical cancellation. The distribution of the false negatives in figure 22c once more attenuates the bias of the classifiers towards $S^{\ominus}$ in the rear of a person. Nevertheless, the coloring of the corresponding observations in the lower left suggests that this mainly concerns formations where other members of a group stood close, but their upper bodies were oriented away from the observer. A similar reasoning applies to part of the observations in the lower right quadrant, but that is also mostly due to increased distance and thus a stronger general bias of the classifier towards $S^{\ominus}$. Vice versa, figure 23 augments the explanation as to what the classifier understands as $S^{\oplus}$ through illustration of the joint distributions of $\delta\theta$, $\delta\varphi$ and δd for the false positives.

2.3.5.4 *Influence of arity*

Considering group size in social situations helps in the investigation of the false negatives. Table 12 shows misclassification rates when arity is taking into account in $S^{\oplus}$. According to these results, the misclassification rate is exceptionally low for social situations of fewer than five participants. With the exception of groups of six, the misclassification rate grows with increasing arity, starting from five participants. Figure 24 explains these results in terms of the joint distributions of the variables for the false positives, where the color of the observations corresponds to arity. In this context, recall that section 2.2.5 made assumptions about the ideal configurations of body orientation ($\delta\theta$) and relative position ($\delta\varphi$, δd) for varying group sizes. Those values that correspond to these ideal configurations have been superimposed onto figures 24a and 24b. It follows that the model is indeed precise for arities two, three and four, resulting in comparatively few false negatives. On the other hand, it also follows that the model is not generally unsuitable for social situations with more participants, but rather for those portions of the corresponding observations that involve greater distances and/or (almost full) frontal configurations. Generally speaking, the distributions of $\delta\theta$, $\delta\varphi$ and δd exhibit the more variance the greater the number of participants. For further reference, appendix B contains illustrations of the joint distributions

Arity	2	3	4	5	6	7	9
# of samples in $S^{\oplus}$	14940	28122	60372	64340	25980	25368	149112
# of false negatives	772	1238	1020	13826	363	5570	59299
Fraction of false negatives	5.2%	4.4%	1.7%	21.5%	1.4%	22.0%	39.8%

Table 12.: Rates of false negatives per group size.



(a)

(b)

(c)

(d)

Figure 24.: Histograms of the joint densities of δθ, δφ and δd, including an orthographic projection of the false negatives per color-coded group size. Diamonds represent the ideal configurations.

per arity. The latter not only sustain this general view, but also give an explanation as to why the classifier performed so much better for groups of six. As a matter of fact, groups of six and seven only rarely showed during the experiment (see table 2), and hence the corresponding observations are not as representative as for other group sizes. Assuming a similar model, it is thus expected that the misclassification rate will actually grow for arities such as six and seven once more data were to be collected.

Other than the corresponding clusters of higher variance, figure 24c reveals a few smaller clusters for e. g. groups of two or four. Interpreting the respective values of $\delta\theta$ and $\delta\varphi$ shows that these clusters belong to rather untypical orientations of the upper body in relation to the relative position. While this figure gives no information on distance, at least the orange clusters for arity two can be quite easily detected in the other figures as well, providing distance and subsequently corroborating the view that involved formations are rather untypical.

2.3.5.5 *The relevance of* $\delta\theta$*,* $\delta\varphi$ *and* δd

The usefulness of each variable for the overall classification task is quantifiable in terms of entropy and mutual information of $\delta\theta$, $\delta\varphi$, δd, as well as the class attribute ($S^{\oplus}$, $S^{\ominus}$). According to information theory, *uncertainty* about the value of a random variable X is expressed in terms of its *entropy* [34, 218], defined as the expected value of the self-information of X:

$$H(X) = \sum_{x} p(x)I(x) = \sum_{x} p(x) \ln \frac{1}{p(x)} = -\sum_{x} p(x) \ln p(x) \tag{93}$$

For continuous variables, *differential entropy* is analogously defined as

$$H(X) = -\int_{x} p(x) \ln p(x) \, . \tag{94}$$

The *conditional entropy* of X given Y measures the remaining uncertainty about X once Y were known:

$$H(X|Y) = \sum_{y} p(y)H(X|Y = y) = H(X, Y) - H(Y) \tag{95}$$

The relevance of variables or features is usually assessed based on the *mutual information* that they share with the class attribute [34, 321]. Mutual information is therefore equivalently referred to as *information gain*. Generally speaking, it quantifies the similarity between the joint distribution of two random variables and the product of their marginal distributions. It is based on relative entropy, also known as Kullback-Leibler divergence:

$$I(X;Y) = \mathbb{KL}\left(p(X,Y)|p(X)p(Y)\right) = \sum_{x} \sum_{y} p(x,y) \ln \frac{p(x,y)}{p(x)p(y)} \, , \tag{96}$$

It follows that $I(X;Y) = 0$ if and only if X and Y are independent [34, 218, 321]. From a set-theoretical perspective, mutual information, joint entropy, and conditional entropy can be seen as set intersection, union, and difference [268], so that the following definition is equivalent to equation (96):

$$\mathrm{I}(X;Y) = \mathrm{H}(X) - \mathrm{H}(X|Y) = \mathrm{H}(Y) - \mathrm{H}(Y|X) = \mathrm{H}(X) + \mathrm{H}(Y) - \mathrm{H}(X,Y)\,. \tag{97}$$

One may note that, whereas entropy of a discrete random variable is strictly non-negative, differential entropy can also take negative values. Its magnitude furthermore depends on the scale of the values of the corresponding random variable. For this, the *uncertainty coefficient*

$$\mathrm{U}(X|Y) = \frac{\mathrm{I}(X;Y)}{\mathrm{H}(X)} = \frac{\mathrm{H}(X) - \mathrm{H}(X|Y)}{\mathrm{H}(X)} \tag{98}$$

provides a normalized measure that quantifies which fraction of X is predictable once Y is known [320]. As a consequence of equation 97, $\mathrm{U}(Y|X)$ is equally easy determined as function of mutual information, regardless of whether the latter was computed in terms of X or Y. Other than that, symmetric uncertainty aids in the quantification of interdependency [348].

For continuous random variables, entropy and mutual information are usually computed based on the previous quantization of the distribution, which may lead to poor results depending on the chosen number and width of the bins, more precisely their exact limits [218]. For the present work, information gain and uncertainty coefficients were hence computed based on GMMs instead of quantization. Moreover, as indicated before, the magnitude of the differential entropy for a scaled variable differs from the value for the same, yet unscaled, variable, i. e. $\mathrm{H}(s \cdot X) \approx \mathrm{H}(X) + \ln s$. The scales of $\delta\theta$, $\delta\varphi$ and δd differ by several magnitudes. As a consequence, all of the variables were normalized (zero-mean, unit standard deviation) prior to computing any information-theoretic quantities.
Recall that the purpose of this analysis is first and foremost the quantification of each variable's importance for this specific modeling task, as opposed to a ranking of the independent features. The dataset was partitioned according to samples belonging to $S^{\oplus}$ or $S^{\ominus}$, and one multivariate GMM was computed for each of these two partitions. The conditional distributions of $\delta\theta$, $\delta\varphi$, and δd were then determined as the marginal distributions of the models for each class, and the total marginals as the respective sums of the conditional distributions weighted by the corresponding class priors. The latter would not be necessary in case of quantized data, but it certainly makes a difference for estimated continuous distributions such as GMMs, or else w.l.o.g. the law of total probability $p(x) = \sum_y p(x|y)p(y)$ would be violated, and consequently falsify the results.

Measure	$\delta\theta$	$\delta\varphi$	δd
Mutual information	0.024	0.047	0.063
Uncertainty coefficient $\mathrm{U}(\mathit{class}|X)$	0.035	0.068	0.092
Symmetric uncertainty $(2 \cdot \frac{\mathrm{I}(X;\,\mathit{class})}{\mathrm{H}(X)+\mathrm{H}(\mathit{class})})$	0.025	0.049	0.062

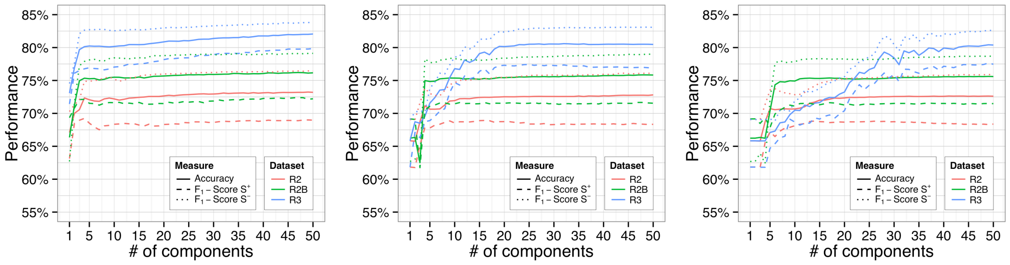
Table 13.: Relevance of $\delta\theta$, $\delta\varphi$, or δd with respect to the class attribute ($S^{\oplus}$, $S^{\ominus}$), given in *nats*.

Table 13 lists the calculated values of the discussed quantities, from which it can be seen that δd clearly has the greatest impact on decisions towards $S^{\oplus}$ or $S^{\ominus}$, followed by $\delta\varphi$. The magnitude of the values attributes to the fact that these variables are by no means independent of the class. Overall, the relation of the values is in line with the previous analysis and interpretation of the data. Whereas δd arguably bears more information due to its higher resolution in terms of the measuring equipment, and while $\delta\theta$ is inherently symmetrical, it is still clear that certain areas of the marginal or joint distributions of specifically δd and $\delta\varphi$ serve as strong discriminant factors. For instance, observations of δd below a threshold of about 50 cm, or in the rear ($\delta\varphi > \pi \bmod 2\pi$), are scarce and therefore very expressive, and so are the distributions of δd and $\delta\varphi$ in comparison to $\delta\theta$.

2.3.5.6 *Summary*

All in all, the chosen model is well-suited for the distinction of $S^{\oplus}$ and $S^{\ominus}$ based on interaction geometry, in particular for social situations involving up to four participants in L-shaped or circular formations. Bigger situations imply configurations which are more difficult to handle as they fall into regions of greater overall uncertainty due to the relations of the actual distributions of $S^{\oplus}$ and $S^{\ominus}$ in the whole dataset. Although the present dataset is of course limited in size, it already captures a reasonable amount of general effects in social interaction geometry. It is expected that these effects become more emphasized also for larger groups as more experiments are conducted and data are sampled. According to the hitherto analysis of the distributions, the interpretation of the false positives, false negatives, the classifier's decision boundaries and how it attempts to explain the data, it is nevertheless likely that such formations impose implicit or explicit limits onto the distinction of $S^{\oplus}$ and $S^{\ominus}$ based solely on geometry of interaction. The misclassification rate is therefore expected to be of at least linear growth with increasing arity. Eventually, this issue will have to be solved by other means. As the model is based on dyadic interaction, and as social situations can be regarded as transitive closure over the latter, mobile agents would therefore have to exchange their opinion and confidence on the who, when and where of social situations.

The overall performance and the characteristics of the model show that context-free interpretation of dyadic interaction geometry is indeed a viable approach for the detection of social situations as a whole. This fact can be regarded as evidence towards the assumption that human behaviour is rather uniform in its very nature, and that it does not necessarily have to be context-sensitive in all of its aspects. To some extent, this uniformity expresses itself in the generalizing way that the classifier explains both presence and absence of social interaction. In regard of the relevance of $\delta\theta$, $\delta\varphi$, and δd, the latter variables apparently provide the most utility for modeling social interaction (see table 13). There is an articulate relationship between the information gain for each of the variables, suggesting that the hitherto model of interaction geometry may be further simplified. This would in turn constitute more evidence towards the uniformity of human proxemic behaviour, similar to the means of the overly simple model of the Naïve Bayes classifier. For comparison, models



(a) (b) (c)

Figure 25.: Comparison of the performances of GMMs (a) and SW-GMMs with one respectively two wraps (b-c) for full and simplified representations of interaction geometry.

were computed based on a representation of interaction geometry solely in terms of $\delta\theta$ and δd (termed “R2”), and yet another representation through $\delta\theta$ and a signed variant of δd, where the sign of δd encodes a binary partition of the space into interaction occuring in front (+) or behind (+) a person (termed “R2B”). The choice of $\delta\theta$ over $\delta\varphi$ is reasonable not only from an information-theoretical perspective but also from the fact that $\delta\varphi$ is much harder to determine from present-day physical and logical sensors of mobile agents. The corresponding performances are shown in figure 25. It follows that a change from the full representation, with body orientation plus relative position given in polar coordinates (as in R3), to body orientation plus only signed distance, therefore losing all information about the polar angle at which another person is located, results in a mere drop of about 5% in performance. Yet another 5% are lost once the sign of δd is dropped, effectively locating the other person on an arbitrary position on a full circle with radius δd around the monitoring person. Note that the simplification of the model also has no additional influence on precision or recall which both drop only about the same percentage as accuracy while maintaining their interrelationship. One last notable difference of R2 and R2B as opposed to R3 is that they need considerably less components when being modeled in terms of SW-GMMs, which is mostly a consequence of the fewer periodic tilings. Performance-wise, though, SW-GMMs again do not exceed GMMs and thus bear no further advantage in this context.

2.4 IMPROVING THE MODEL THROUGH ADDITIONAL PARAMETERS

The proposed model so far has been developed under the hypothesis that universal behavioural patterns exist in proxemics, and hitherto evaluation corroborates this assumption. On the other hand, it has been noted that the experimental dataset is arguably limited in size, both in terms of the actual data as well as in its representational sense, being based on measurements of young adults from Western Europe, only two of which were female, some acquainted via university, and all coming from a more or less equal cultural background. It is well worth questioning whether such (and other) parameters do in fact have a significant influence on social interaction geometry – and if so, to what extent. That goes along with the question as to how the presented model could and should be adapted to any corresponding findings, for example by incorporating a subset of personal profile parameters. Without doubt, there is a multitude of possible parameters like age, gender, culture, physical appearance, health, profession, and social background, to name only a few. In addition to that there are also other (latent) variables like e. g. the number of interactants, environmental or other situative attributes that one might want to consider to incorporate under the objective of exhaustively modeling social interaction. Vice versa, a particular model which incorporates some of these (latent) variables could possibly provide information about their specific values and weights when applied to a new data. For example, think of a function $p(n|x, \theta)$ yielding the probability of participating in an n-ary social situation, given observations from interaction geometry and the respective model parameters. Another example could be a function from group size to the presumed character of any dyadic relationship in that group, be it casual, talking to a superior or potentially participating in a focussed discussion.

The following sections investigate the influence of a selected subset of profile parameters, as well as other (latent) variables, and how they could be exploited or otherwise be used to improve the present model.

2.4.1 *Influence of profile parameters*

A common sentiment in literature is that behaviour is mostly a consequence of factors such as the environment, the affective meaning of a situation, the behaviour of others, emotions, and that it is strongly influenced by personality traits [315, 325, 198]. Already Hall was aware that “social organization is a factor in personal distance” [133]. The fact that part of proxemic behaviour is learnt early during socialization implies a high correlation with the culture of the society in which individuals grew up [133]. Behaviour can be regarded as a function of socio-cultural background [250]. Proxemic behaviour generally happens unconsciously [131] and is usually beyond a person’s locus of control. The understanding and interpretation of situations, and consequently one’s behaviour, strongly depend on individual personality. For this reason every situation has a so-called *affective meaning* [286]. According to [319] in [286], “as individual people strive for emotionally coherent mental representation, they mutually coordinate their social actions – both verbal and non-verbal – so as to make them fit those representations.” People hence try to work towards, and subsequently maintain, the affective meaning of a situation. In this regard, *communicating* affective meaning plays an important role, and people consciously or unconsciously act together in order to achieve a collective affective meaning, expressed in terms of verbal and non-verbal communication [286, 250, 198]. Triandis further explains this effort by differentiating between the private, public and collective selves [325]. The private self accounts for personal attributes, preferences and one’s individual character. The public self aims at acting in compliance with what others would deem appropriate, and retrieving corresponding feedback. The collective self pursues the common goals of the peer group. It follows that whereas all participants of a social situation have their share in the display and outcome of behaviour, they still each have their own motivations and objectives.

The above belongs to a vast field in socio-psychological research. Following the seminal works of Hediger [140] and Hall [130, 131, 133] (see chapter 1), a multitude of studies have been conducted by anthropologists, psychologists, sociologists, and many others. These works typically investigated the influence of culture, gender, age, attractiveness, or friendship and acquaintance [309, 110]. In what follows, an overview shall be given over concerns about culture, gender, and a subset of the remainders, as they have been the main focus of related research, in that order.

2.4.1.1 *Culture*

In [132], Hall labels proxemics as a “culturally elaborated form of communication” . This view is based on the notion that “language is a major element in the formation of thought” [22]. Amongst other things, culture expresses itself through non-verbal communication. The subsumption of any form of communication under the term language, together with the assumption that language has significant influence on the process of thought, leads to the conclusion that cultural aspects have a particular impact on even basic sensoric perception, and therefore also on the assessment and interpretation of proxemics.

Hall gives several examples of cultural effects. In regard of perception and well-being, for instance, he explains how a frontal seating arrangment might be perfectly normal for Europeans whereas in China there might be a connotation with being on trial [131]. In yet another example, Hall refers to an Arab colleague of his whose paneled recreational room felt cozy to his German friends yet oppressive to his Arab friends [132]. Other than that, cultural differences can also result in sociopetal and sociofugal effects, as e. g. furniture is not only organized differently in Europe as opposed to Japan, where Europeans place their furniture at or near the walls, whereas Japanese are apt to clustering everything in the center of a room [133], but as well influences the duration of social interactions [166]. Much like Hediger made a difference between contact and non-contact species [140], Hall differentiates between supposed contact and non-contact cultures. He assumes that members of contact cultures are more likely to stand closer together, talk louder, touch more often, and be more directly oriented towards each other during interaction. As examples he names Americans and Arabs for contact and non-contact groups, respectively [131, 133]. Hall furthermore states that e. g. Westeners from non-contact groups might even associate crowdedness with “distasteful connotations”, therefore attenuating his belief in the strong influence of culture. He goes as far as saying that erroneous behaviour in contrast to what is acceptable within one’s culture can lead from dissent to even real anxiety [132]. Others share his view in that all members of a culture are supposed to behave accordingly, and that consistency of verbal and non-verbal interaction allows for mutual understanding [97, 286]. Culture bears latent core knowledge about successful behaviour throughout numerous generations in the past, enabling individuals to “derive intuitive knowledge about the probability and cultural appropriateness of mutual behavior, including non-verbal expressions of interpersonal affect.” [286].

One should note that, although Hall states that a lot of his observations are based on actual fieldwork, his theories on contact and non-contact cultures as well as other cultural influences were largely speculative. Watson et al. therefore set out for a quantitative investigation [340]. They confirmed that groups of Arabs indeed communicated at closer range and with a louder voice than Americans. They also confronted each other more directly in terms of relative position and upper body orientation. Arabs had a tendency to (accidentally) touch each other, whereas Americans avoided touching at all times. As touching occurred in all of the Arab but none of the American groups, Watson et al. conclude that this may be an outcome of culture. It also turned out that the variance in behaviour was way less among different groups of Americans from different locations in the United States than of Arabs from varying origins. At the bottom line, their findings seem to sustain common stereotypes. However, the accuracy of the annotations was rather low, and only thirty-two individuals were monitored during the process. Interestingly enough, Watson et al. were the first to state the question if variations in proxemic style *within* cultural areas are perhaps associated with other personality traits. In a later study, they were furthermore able to show that the proxemic behaviour of individuals from supposed contact cultures adapts itself over time spent in non-contact cultures [339]. Little [193] conducted yet another study in order to find differences with respect to interpersonal distance among Americans, Swedes, Greeks, Italians, and Scots. Based on the placement of dolls and subsequent assessment by experimental subjects he concluded that there were differences between all of these cultures, most significantly when compared to the Italians. In spite of the fact that his results support the distinction of contact and non-contact cultures in principle, he also found that there were less differences between Americans and Italians than anticipated, which is remarkable as they are supposed members of contact and non-contact cultures, respectively. Shuter likewise investigated the proxemic behaviour of Germans, Italians, and Americans [301], more precisely the frequency with which interactions occurred, interpersonal distance, relative orientation, and gender. According to Schuter, the stereotypical distinction of contact and non-contact cultures is not sufficient because of the high variance in intra-cultural behaviour. His results for example show that the overall observed interpersonal distance is greater for Americans than for Italians, yet in terms of physical contact or touching, there is no apparent difference between Americans and Italians for male-female and female-female dyads. Quite to the contrary, physical contact was observed between German even more than between Italian females.

Apart from culture, Little also looked into variables like gender and social roles such as e. g. dominance or authority. There was no apparent general difference due to gender, so finally the major determinant was attributed to the relationship between individuals within a group, followed by the affective tone of the transaction.

In a field study of 859 subjects in “several natural settings” over the course of 2 months, Baxter [28] found even more statistical evidence for Hall’s and Little’s assumptions. He observed “Anglo-, Black-, and Mexican-American” ethnic groups, and noted other factors like age (in three discrete levels, i. e. child, adolescent, adult), as well as the gender of dyads. According to his results, there was a striking "tendency for Mexican subjects of all ages and sex groupings to interact most proximally", which he regards as members of a supposed contact culture, as opposed to the other ethnic groups. The differences in interpersonal distance were already clearly apparent for children, and increased with age, suggesting that spatial schema are learnt at young age and similarly retained through adulthood. Motivated by the awareness of the effort of groups to maintain formations and to compensate behaviour of others in this process, e. g. if one person decreases interpersonal distance and others consequently take a step back, Baxter suggests that ethnic groups be mixed in further experiments. Assuming real differences in the proxemic behaviour of different ethnic groups, this would show in that members of respective groups would work towards different goals [28]. In a 1970s paper [189], Leibman assumes the existence of measureable effects for interpersonal distance due to ethnic differences [189], and discusses that this might as well be a consequence of the culture among "white" and "black" Americans. She further states that interaction always depends on context, but from her paper it remains unclear which variables are the ones which have an actual influence on behaviour. Her experiments, however, showed no significant results, but "indicate that the social environment is a significant determinant of the perception and use of space, and that spatial behavior is an important measure of the behavioral consequences of social factors" [189]. In yet another study, Jones [157] observed pairs of interacting persons from several of New York's "subcultures" within previously determined and strictly defined regions, and at places of equivalent *socio-economical account*. In addition to interpersonal distance, Jones also noted relative orientation of the dyad's shoulder lines. His results vaguely augment those of Leibman in so far as he did not detect statistically significant differences in either interpersonal distances or relative orientation between members of distinct subcultures. It should be noted though that his methods and results are questionable because the measurements where subjective and imprecise, not least because the subjects were observed from a certain distance, and interpersonal distance as well as orientation were only roughly estimated.

Although most of the presented studies seem rather inconclusive, they do indicate cultural differences in proxemic behaviour. All the same, it is not clear whether culture alone, or ethnic group, or heritage, or some other yet unknown variables account for this. This is corroborated by Remland et al. [265], who argue that the principal influence of culture might be either less than hitherto anticipated, or that differences are more likely to come from latent variables such as social relationship, emotion, and personality traits [265]. Sussman and Rosenfeld, for instance, showed in a study of Japanese, Venezuelans, and Americans that how close people sit to each other is influenced by the language which is spoken, be it native or learnt, which they assume as a consequence of aiming at the display of appropriate behaviour when concentrating on another culture along with speaking a foreign tongue [316]. This is sustained by a later study of Remland et al. on interpersonal distance, body orientation and touch between North- and South-Europeans, evaluated according to either origin, gender, or age [266]. The results demonstrate differences in physical contact behaviour, and likewise body orientation for mixed male-female and male-male dyads with respect to age, but also reveals that these are neither related to contact/non-contact cultures nor any other “generalizable function” from North to South. Similarly, Evans et al. [91] discuss that cultural background has often been mistaken for a tolerance of crowding when instead this has been related to personality. Amongst other variables, according to Sommer, the dimensions of personal space are particularly dependent on culture, “internal state”, and transactional context [309].

2.4.1.2 *Gender*

Most of the previously mentioned studies focussed on cultural influences on proxemics, although some of them also reported findings with respect to gender [307, 28, 157, 301, 266]. Women arguably stand closer together than men, adopt more direct orientations, whereas men are less apt to physical contact during interactions. These presumptions are to some extent corroborated by socio-psychological studies which quantify and evaluate personality traits based on ratings of the Five Factor Model [77]. The model organizes personality traits in equivalence classes of *extraversion*, *agreeableness*, *conscentiousness*, *neuroticism*, and *openness* [208]. Interestingly, women are believed to score higher on all factors except openness [203]. Extraversion, agreeableness, and conscentiousness are particularly interesting. Roughly speaking, extraversion represents sociability and positive affect, agreeableness stands for trust, warmth and kindness, and conscentiousness describes self-control, task orientation and rule abiding [203]. Together, extraversion and agreeableness may be linked to smaller interpersonal distances and the allowance or even initiation of physical contact, especially touching. Between males and females, the scores for conscentiousness vary less than in other categories. Conscentiousness may however help to explain observations like those of Hartnett et al. [137] who reported that, under given circumstances, women would let male experimenters approach up to closer distances than would men.

Jones reports that observed female-female dyads were generally more directly oriented towards each other in terms of their shoulder lines than others [157]. This is sustained in part by Shuter who observed that American male-male dyads interacted at a significantly less direct axis than male-female or female-female dyads. On the other hand, German subjects demonstrated the opposite behaviour, namely that in male-male dyads the relative orientation was the most direct [301]. Likewise, in his early studies of seating arrangements, Sommer observed that men were more likely to sit in opposite chairs to either men or women, while women had the tendency to sit next to each other, or at adjacent corners of a table [307]. To the contrary, Remland et al. [266] found that male-male dyads maintained the least direct orientations until an age of about sixty years, and that orientation becomes more direct with age. Mixed dyads confronted each other most directly until about forty years, and orientation would lessen with age. Lastly, for female-female dyads, orientation would stay roughly the same at each age, with the exception of forty to fifty years where women would surprisingly adopt the most direct orientation of all [266]. These relatively recent findings therefore contradict the prior opinion that women would generally adopt more direct body orientations than men.

Apart from body orientation, Shuter [301] reports that women were more likely to touch other women than men were likely to touch other men. The least amount of touching was observed in mixed dyads. Shuter also distinguished between touching and hand-holding. Most notably, hand-holding was regularly observed among females, regardless of nationality, whereas for male or mixed-sex dyads, most hand-holding was observed for members of supposed contact-cultures, such as Italians. According to Berman and Smith [31], "less attention has been paid to the implications of touch between participants of equal status as a sign of friendship, support, and solidarity". In their exploration of 256 subjects, no significant differences were found for touching or proxemic behaviour between genders. Berman and Smith therefore account touch and proxemics solely to the type of social situation. Similar results were reported by [266]. Touching occurred most often in mixed, but equally often in same-sex dyads. This was consistent throughout all groups for non-hand touches, or touches that lasted longer than 2 seconds, for instance when touching or holding someone else's arm. According to [216] in [266], such behaviour is typically related to a "tie sign" between partners. As touching is often a reciprocal reaction to being touched, it is argued that *general* behavioural differences between males and females should be investigated after successful determination of who initiated the contact.

In a study of 186 subjects, all of which were introductory students of psychology, Dosey and Misels investigated the influence of stress and gender on interpersonal distance [78]. Pairs of the same, opposite or mixed sex were randomly chosen and the approaching distance was measured, i. e. the distance until which one would approach the other and then stop on their own. The results showed that women would approach other women more closely than they would men. From the perspective of men, there was no notable difference whether men would approach other men or women. Although most variance in interpersonal distance in the previously discussed work of Baxter (see section 2.4.1.1) was accounted for by culture [28], a significant part was also reported due to gender [28]. In accordance with later studies, male-male dyads demonstrated the greatest distances, whereas mixed dyads showed the least distance. This is somewhat contrary to the notion of female-female dyads as the most closest interactants. At the bottom line, Baxter states that it is not exactly clear whether his findings were generally due to gender or culture, as presumed friendly or familiar relationships were more often observed in certain ethnic groups than in others. In a similar laboratory study, Hartnett et al. [137] further distinguished between approaching distance and distance when being approached. In order to eliminate any territorial, social or cultural effects, experimenters wore white labcoats and were instructed to act without and show no emotion. In addition to influences caused by gender per se, the study also evaluated the heterosexuality score of each participant so as to determine whether persons with a high heterosexuality score, and thus a supposed higher interest in the opposite sex, would exhibit a tendency towards smaller distances. The results [137] show a notable trend, though not significant, where men with a higher score would let women approach up to closer distances. In general, women would not actively approach other persons or experimenters as close as they would allow for being passively approached, especially by experimenters. It is discussed whether this may be a consequence of social norms with respect to the gender role of females, or generally the “aggression behaviour” of men versus women, i. e. that women perhaps demonstrate slightly more obedience under “official circumstances”, such as a laboratory experiment. If this were true, it would agree with the finding that women tend to higher scores on agreeableness (being task oriented and rule abiding). To the contrary, this may as well not be linked to gender at all, but merely be an effect of respect and/or social state, as e. g. reported by Cristani et al. [67].

According to Uzzell et al. [330], the present-day results of researching the influence of gender on interpersonal distance are ambigious. Nevertheless, there seems to be a certain agreement that, generally speaking, male-male dyads interact at greater distances than pairs of females. This is sustained by the studies of Evans and Shuter [90, 301], and may be further explained in terms of territoriality and personal zones. Edney, for instance, observed groups of people on a public beach [85]. He found that groups of three (male, female, or mixed), as well as groups of only females, would occupy considerably less space than groups of four (male, female, mixed), any group of men, or single men. Interestingly, mixed groups used less space than groups of either men or women. The authors suppose other influences like the actual reason for visiting the beach, like being there with one’s family, friends, or as a single. Following their measurements on the so-called comfortable interpersonal distance scale, Veitch et al. [332] conclude that the personal zones of men and women differ indeed. Women generally have smaller personal zones ($\sim 2.32m^2$) than men ($\sim 2.79m^2$). The same relation holds for the accumulated zones of female-female versus male-male dyads.

As mentioned before, Hartnett et al. were among the first to account for more than gender per se when they tried to quantify sexual attraction through a heterosexuality score during their assessments of approaching distances [137]. Likewise, Uzzell et al. suggest the clear distinction of *biological sex*, *gender role*, and *gender identity* [330], where gender role “is a label for the masculinity or femininity of someone’s (social) behavior”, and gender identity refers to the psychological sensation of the actual biological sex, which might a factor e. g. for transsexuals. They relate to previous studies of [139] and [293], according to which gender alone has no sufficient meaning, but has to be seen in conjunction with race and age. In a similar manner, West and Zimmerman made a distinction between sex and gender [346]. In their quantitative study of 72 participants, for which they used measurements from digital video recordings, they ultimately concluded that gender role is in fact responsible for more variance than biological sex [330]. Others presume that the behaviour of men and women is first and foremost affected by their way of self-representation, and hence suggest that gender-specific differences always ought to be interpreted in the social context in which they occur [31]. Ridgeway and Smith-Loving thus postulate in [270] that all theories regarding the influence of gender should respect three basic aspects: First, women and men alike perceive gender as a profound factor in interaction. Second, studies of women and men with equal power or state generally fail to demonstrate significant differences. Most differences are probably due to socio-emotional, non-verbal contexts which are likely connected with expressed or displayed behaviour. Third, most interactions between men and women take place in a structural context which already implies different roles or status. Correspondingly observed differences are therefore likely to be confused with being related to gender.

2.4.1.3 *Other parameters*

A couple of studies have explored the influence of further profile parameters and variables. Reis et al. e. g. supposed that *physical attractiveness* has potential influence on how humans interact in social situations [263, 264]. According to their results, this factor would mostly affect men. More precisely, physically attractive men tend to more social interactions with women than with other men, while attractiveness in general attributes positively to the affective quality of social experience for both genders. Next, Dosey et al. describe personal zone as a buffer zone whose main purpose is the protection of *emotional well-being* [78], which, amongst other things, is a function of *stress perception.* Arguably, the way that stress is perceived can be thought of as a personal parameter. In their study, Dosey et al. monitored the interpersonal distances of 189 subjects, partially under induced stress. They found that distance significantly increased under stress, and report an average distance increase from 6.35cm to 9.5cm. Other than that, according to Edney [85], it appears as if *occupation* had an effect on territoriality and hence the personal zone. Moreover, Cook investigated whether being *introverted* or *extroverted* influenced proxemics in terms of seating behaviour [61]. According to his results, extroverted persons have a tendency of moving in or sitting closer, and are apt to adopt rather frontal configurations. The behaviour of extroverted persons was also reported to be the most consistent. His view is shared by Patterson and Sechrest who found that the personal zones of extroverted persons are smaller than those of introverted persons [227]. In addition to his findings that gender may attribute to the choice of seating arrangement and distance, Sommer also explored whether *mental health* (of both sexes alike) might be influential. He assumed that due to the fact that mentally ill people often have problems in their communication with others, their problems might alter their proxemic behaviour. For this, he observed both patients and healthy persons in a mental hospital [307], and found that schizophrene persons tended to sit closer to other persons. This tendency towards violating the personal space of others is contradictory to Hall who reported that schizophrene persons would often describe violations of their own space as a feeling of the other persons being literally inside them [133]. Likewise, Evans [90] argues that individuals with personality disorders or other mental illnesses need more space than others.

Another class of studies reports that proxemic behaviour depends on *age.* Both Heshka [143] and Baxter [28] found that younger and elder dyads stood closer together than middle-aged ones. The tenor is that distance increases with age. Starting from an age of about forty years, this appears to be counteracted by other effects such as loss of hearing or sight [143, 28]. As discussed before, Remland et al. [266] report a high correlation between age and body orientation, particularly so for women between forty and fifty years who adopted the most direct body orientations towards others in comparison to all other groups of age or gender. Marsh et al. [203] refer to an exhaustive and complex study by

Terracciano et al. [318], according to which neuroticism decreased non-linearly with age, and so did extraversion, whereas a linear descent was found for openness, as opposed to a linear ascent in agreeableness. Conscientiousness was reported to grow until an age of about sixty years, followed by a subsequent decline (both non-linear). Age thus certainly has an impact on social behaviour and proxemics, but it remains unclear whether the influence of age could be canceled out as an independent variable.

Without doubt, *social relationship* is a major determinant in proxemic behaviour, and so are *topic*, *purpose* and "*tone*" of a transaction [193]. Although equal status may often be anticipated, in many encounters "some participants have different rights than others" [167] and "this, too, is reflected in spatial-orientational arrangements" [167]. Sommer [306] already mentioned that e. g. the seating arrangement of individuals is a consequence of purpose. People who want to work together sit next to each other, those who want to chat tend to sit at adjacent corners of a table, rivals choose opposing places, and strangers likely maximize distance. Sundstrom and Altman [315] describe what they call the comfortable distance, which varies with the status of a relationship and also depends on additional parameters such as whether people sit or stand, the topic of a discussion, gender, orientation, and crowding [132, 133, 315]. The dependency of interpersonal distance on type and quality of relationship was discussed by Bell [29]. Heshka et al. observed interpersonal distance of 57 subjects under the influence of whether they were good friends, acquaintances, or strangers [143]. They later combined the first two categories as it turned out that too often one would assess the other as a good friend while being assessed as an acquaintance. No significant differences were found for male-male pairs, but instead for female-female and mixed dyads. Not quite unexpected, the behaviour of male strangers differed much from female and mixed stranger dyads. Women positioned themselves significantly farther apart from others than men, which Heshka et al. relate to the socializing process "which encourages aggressiveness and initiative in males, and caution and reverse in females" [143]. On a sidenote, interpersonal distance between stranger male-male dyads was reported to be less than as assumed by Hall for American same-sex strangers [129]. Also recall the finding of Cristani et al. [67] according to which physical distance is proportional to social distance, and hence interpersonal distance is a function of mutual relationship. Vice versa, distance can hence serve as a social cue for the type and quality of a relationship.

2.4.1.4 *Critique*

The previous sections have outlined related research on potential influences of profile parameters like culture, gender, age, and other variables. It appears that a number of these studies are inconclusive or even contradict each other. Some studies have measured significant influences which were not reproducible by others. Arguably, it is presumed that this is more often than not a consequence of the fact that other latent variables may be involved which are yet unknown. Jan et al. [153] remark that Hall not only failed to provide any form of qualitative or quantitative proof for his theories regarding e. g. the influence of culture, but also that subsequent studies were often not exhaustive or subject to systematical or methodical errors. Some studies explored effects in the context of ongoing interactions, yet others deduced their results from measurements in the context of personal space invasion [315]. The three prevalent research methods in the field were identified as simulation methods, laboratory methods, and field methods [315]. Simulation methods, for instance, frequently make use of doll, figure or symbol placements, followed by subsequent assessment by the investigated subjects. In laboratory settings, subjects also knew they were being observed, and (w.l.o.g.) controlled settings as well as controlled (latent) variables are often hard to guarantee. Lastly, field methods are based on observations in everyday settings. This is likely to have a positive effect on canceling out other variables and on the subjects' behaviour who are typically not aware of the fact that they are being monitored. Field methods can however have adverse effects, namely that unknown variables or specific portions of the settings have a potential influence on the results. Also, earlier studies using field methods were particularly prone to inaccurate measurements and subjective assessment of the experimenters, for example the studies on interpersonal distance and relative orientation in selected New York subcultures, [157], which were assessed from a non-negligible distance and by means of a rule of thumb, or those of dyads in natural settings such as parks or public places [143], for which the measurements were taken from distant photographs [157, 143].

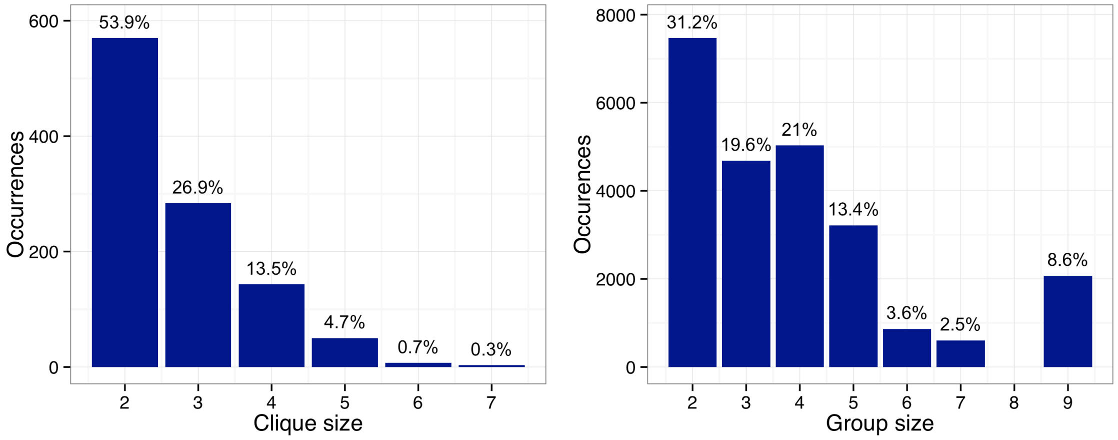
Research nonetheless agrees that Hall's theories have yet to be refuted. In fact, despite all criticism it seems almost certain that there are significant influential factors on proxemic behaviour, and researchers emphasize the importance of conducting more studies to achieve definite results [153]. This would hopefully overcome reliance on overly simplified distinctions such as contact and non-contact cultures, or gender in terms of biological sex. Remland et al. [265] further emphasize that most related research has been conducted in America. Doing more research in other countries (or, generally speaking, in other contexts) could only help to determine what could possibly be regarded as the greatest common denominator. It should be noted, though, that some researchers were well aware of the typical shortcomings in their research, such as e. g. Watson et al. [339], who stated their need for better technical equipment and methods for more accurate measurements and improved quantative studies. Most notably, they were furthermore interested in the ability to work on smaller spatio-temporal scales. Other researchers similarly express the need for methods which are not solely based on pure human observation or manual video analysis but are supported by more sophisticated technical means [265]. As discussed in chapter 1, this would allow subjects to act more freely and unconsciously, or achieve higher accuracy and resolution [137]. Nowadays, the high advancements in sensors and fields such as Computer Vision have greatly alleviated issues of accuracy, precision, and smaller spatio-temporal scales. Corresponding techniques have been used in laboratory and/or field settings e. g. by Groh et al. [123] or Cristani et al. [66].

2.4.1.5 *Arity*

Common shapes of FFSs include L-shaped, V-shaped, side-by-side, circular, semi-circular, rectangular, or linear formations [204, 166]. L-shaped, V-shaped and side-by-side formations are common representatives of pairwise interaction, whereas three persons often adopt a triangular (circular) or semi-circular formation, and greater groups exhibit a tendency towards circular formations of increasing size. Territory grows with the number of persons, but its growth is not regular because the space which is occupied by a single person appears to be inversely proportional to group size [85]. Marshall et al. [204] also differentiate between FFSs and what they call a *common-focus gathering*. As an example for the latter, consider a group during a museum visit. A member of that group might temporarily leave the group and shortly after return from an information desk to share the obtained information. Such common-focus gatherings are more closely related to audience situations than to regular FFSs [204], which also follows from the definition of FFSs, as under the given circumstances not all members have equal access to O-space. Next, recall that formations are usually chosen and communicated unconsciously, and that groups work together in a natural effort to create and maintain a common spatial-orientational configuration [167]. Which formation is chosen eventually depends on the whole context of the interaction, which to a large degree is composed of physical and socio-psychological aspects. Thus a formation may be e. g. chosen as a result of the geometry of the environment, potential obstacles, or it may be subject to sociofugal and sociopetal effects such as those imposed by furniture or architectural design. Socio-psychological aspects may attribute to the arrangement e. g. in terms of social relationships or the purpose of the transaction, for instance when interacting with a close friend as opposed to a superior at work, riding on a subway, being a member of an audience, or sharing a table at a restaurant. However, the problem of gaining continuous and reliable access to accurate, current and exhaustive information about the physical and social context is intractable, especially in a mobile computing scenario. Despite the numerous potential physical and socio-psychological aspects, one major determinant in the choice of formation is the number of interacting persons. Although by far not exhaustive, proxemic behaviour certainly is also a function of group size, in terms of the occupation of space as well as the arrangement of the individual bodies can be regarded as a function of arity. Arity, as opposed to other (latent) variables, is *quantifiable* and can be measured *independently* of other contextual parameters. It is also easily monitored or controlled as an experimental parameter.

Conversational groups are never unlimited in size [79]. The higher the number of participants, the more likely it is that subgroups split off permanently or temporarily, and/or regroup at later times [206, 79]. This behaviour can also be observed in the present dataset (see section 2.2.3), although that may not be strictly comparable. Recall that the subjects were given instructions which would foster that they would regularly switch between, or form new, groups. The reason behind subgroups splitting off from larger groups may be related to the presumption that group size is limited due to cognitive abilities of humans [80]. According to Hall [133], the quality of sensory input is inversely proportional to distance. This also means that the greater the distance, the more social cues have to be taken into account to obtain the same amount of information. Eye-sight, for example, influences group size. As the field of vision is limited, humans can track only a certain number of others, and maintaining at least peripheral view on other interactants is a vital component of groups [79]. According to Kendon [166], humans have a field of vision of about 80° to either side before having to turn their upper bodies. Assuming an ideal circular configuration and a distance of 70cm between adjacent persons (see section 2.2.5) therefore leads to a theoretical limit of eleven persons. Other than that, recall that FFSs require a significant overlap of the transactional segments, and therefore also imply a limit on the number of interacting persons. Dunbar [79] furthermore distinguishes between *groups* and *cliques*, for which group size refers to "the total number of individuals present in an interacting group" and clique size denotes the "the number of individuals taking part in a particular conversation, as evidenced by speaking or obviously attending to the speaker" [79]. Cliques seemingly obey a natural limit of four persons, independent of gender. The reason for such a practical limit might be rooted in the production and detection of speech. According to [305] in [79], the maximum comfortable distance for conversation is 1.7m. Dunbar concludes that this "imposes a limit of five on the number of individualds who could take part in a conversation", given a respective distance of 50cm between adjacent persons in circular arrangement. Comparing this to the present dataset, according to which 70cm are closer to the truth, one may hence estimate a maximum of $2\pi r/0.7m$/person $\approx 7 persons$ for $r = \frac{1.7}{2}m$. This estimate would then also be in agreement with Cohen [57], who came to an equivalent conclusion, albeit under consideration of ambient noise.

A quantitative study with 1057 groups of up to fourteen individuals is described in [79]. The major part of the groups were observed in a college dining hall. Groups were sampled every fifteen minutes as long as all members remained. In addition to the dining hall scenario, groups were also observed in the context of casual talks after a firedrill, as well as during a large reception at a museum. Generally speaking, it certainly makes sense to make a distinction between *cliques*, as interacting subsets of a group, and *groups* as a whole. In regard of Dunbar's study, it can however be assumed that the environment of the dining hall, more precisely the seating and table arrangement (Dunbar describes tables that could serve up to thirty persons), might have had a non-negligible influence on the formation of cliques within larger groups. To some extent, the same may be true for the museum reception. As a matter of fact, Dunbar's paper contains no precise description of that environment, but one may argue that such events often feature numerous smaller tables which clearly yield sociofugal and sociopetal effects. Therefore it is assumed that cliques respective groups could not freely adopt FFSs during Dunbar's experiments. In regard of the present work, and ignoring the discussed spatial constraints due to the infrared tracking process, groups could freely move and split at all times. Arguably, not all members of the groups were active interactants at all times, but cliques could easily split off and form new groups in their own right. A clique in Dunbar's sense is hence more closely related to a group in the present dataset, whereas a group in Dunbar's sense is more closely related to all participants of the experiment as a whole. This is further corroborated



(a)

(b)

Figure 26.: Distribution of cliques as reported by [79] (a) vs. groups from the present dataset (b).

by the finding that clique size grows about linearly for groups of up to seven individuals [79]. Interestingly enough, Dunbar observed that maximum clique sizes were larger for groups of six to nine than in even bigger groups, which he relates to an "overshoot effect, in which individuals initially try to maintain the group as one clique as its size increases" [79].

Figure 26a illustrates the distribution of clique sizes as observed by Dunbar [79]. Similar distributions were reported by other researchers [152, 217] for freely forming groups in various settings. With 15486 respective 1353 sampled groups, the latter seem to be equally reliable. Also, since both studies investigated groups as a whole, not cliques, the reported distributions further attribute to the argument that groups in the present work relate to cliques as seen by Dunbar. For comparison, figure 26b shows the distribution of group sizes from the present experiment. The maximum clique size as observed by Dunbar is seven, whereas up to nine individuals formed a group in the present experiment. The latter is presumably an effect of the rectangular space of the recording area. According to table 2 and figure 5, groups of nine actually showed up twice during the experiments. The observations each cover relatively short time spans, yet serve to illustrate the tendency of larger groups to split up temporarily and then regroup. All in all, both distributions in figure 26 are roughly similar. It should be noted that in spite of the much greater number of occurrences of groups during the present experiment, the actual total number of groups (34) is less than Dunbar's (1057) by almost two orders of magnitude (see table 2). Dunbar mostly assessed groups at time frames of fifteen minutes, whereas time frames correspond to only one sixth of a second in this work. It is therefore most likely that the distribution of groups, as present in interaction geometry, will approximate Dunbar's distribution along with increasing numbers of samples. At the bottom line, the similarities of the distributions, together with the much larger sample size in the aforementioned studies, suggests that

Dunbar's distribution may be tentatively regarded as an appropriate *class prior* for group cardinality once. As such, it could easily be incorporated in the proposed algorithmic model for detection of social interaction.

2.4.1.6 *Discussion*

The prior sections discussed potential influences of profile parameters and (latent) variables on proxemic behaviour. So far, related research has mostly investigated factors like culture, gender and age. In addition to such *a priori* available personal profile parameters, group size was discussed as another major determinant in interaction geometry. Practically speaking, the influence of the latter was already apparent from the marginal and joint distributions of $\delta\theta$, $\delta\varphi$ and δd in the experimental dataset (see section 2.2.5; figures 9, 10, 11, 24; appendix B). Up to this point, the proposed model has been built on the assumption that a greatest common denominator for interaction geometry exists regardless of other parameters. A corresponding classifier has been evaluated and the results indicate that this assumption is indeed valid, if only up to a certain degree. Subsequent analysis of the misclassified samples, in particular the false negatives, suggests that using distinct and more specialized models, e. g. conditioned on group cardinality, might improve performance. On the other hand, one has seen that as group size increases, so does variance, consequently leading to poor classification performance, in particular for groups of five or more (see table 12). This finding is further corroborated by qualitative comparison of the distributions of the variables for $S^{\oplus}$ and $S^{\ominus}$, independent of group size (see figures 7).

In order to determine whether incorporating such kind of parameters into the model has indeed impact on the detection of social interactions, and, vice versa, whether doing so allows for gathering *a posteriori* information such as group size, a new dataset was sampled and adaptions were made to the model accordingly. From the set of potential parameters, gender and arity were selected as representative variables for the following reasons: As the number and domains of potential variables are practically unknown, comprehending and integrating all of them is pratically impossible, and so is controlling all variables in either field or laboratory settings. Furthermore, once the number of independent variables increases, the number of required samples grows exponentially [34]. Gender and arity stand out because both are quantifiable and controllable. Culture, on the other hand, is rather a superposition of numerous known and unknown factors. It is presumed that culture cannot be measured on a discrete scale, or may even be subject to changes after spending a certain time abroad, regardless of the original culture of an individual [339]. Following the findings and suggestions of [137, 346, 330], it is still arguable whether gender should be regarded in terms of biological sex, or gender role, or gender identity. Also, recall that gender-related experiments should be conducted under circumstances of equal power and status, so as to rule out other factors from a socio-emotional, non-verbal or structural context [270]. Distinguishing between roles would however require a substantial increase in the number of subjects. Nevertheless, it is assumed that the overall distinction of biological sex as a dichotomous variable is sufficient for the determination of basic influences of gender on proxemics, even if it will remain unclear precisely which one of the three subcategories had the most impact. Apart from gender, recall that arity was deliberately not controlled during the first experiment, but would be easy to control and monitor throughout subsequent experiments. From the previous discussions in section 2.4.1.5) it follows that groups of more than four participants are relatively unlikely. This, together with the fact that the distributions of the variables from the first experiment show minor effects due to the restricted size of the available space in which interactants could freely move (see section 2.2.5), suggests a maximum group size of four for subsequent experiments under the given conditions.

2.4.2 *A second dataset*

As a consequence of the considerations layed out in section 2.4.1.6, another series of experiments was conducted in July 2013 [81]. In this series, groups of two, three, or four participants were observed over the course of about 15 minutes each. All experiments were conducted and therefore all groups observed individually. Also, the participants of the groups were selected such that groups were composed of either males only, females only, or both sexes. The small group sizes were a deliberate choice to ensure that all individuals were part of a single social situation at all times, and that they could freely move about in the available space of $3\mathtt{m} \times 3\mathtt{m}$. It was explained to the participants that the experiments were about algorithmic models of social interaction, but no further details were provided so as to reduce the risk of additional influences due to the laboratory setting. The participants were asked to engage in casual conversation. Example topics like someone's "favorite meal" or "best vacation" were printed on posters and distributed throughout the room, intended to prevent conversations or interaction from coming to a halt. Each session was monitored by two experimenters who took great care to not engage in any interaction with the participants and displayed behaviour as if they were occupied with doing other things aside from the experiment without even noticing what was going on. All the same, the experimenters took notes on the general atmosphere and whether the groups were actively involved in conversation. According to the results in [81], groups "quickly found a subject of conversation that appealed to all members", regardless of the suggested topics, and "awkward pauses never occurred at all.".

Over the course of three weeks, 30 males and 21 females participated in the experiments, most of which were students, whereas others were not affiliated with university at all. In advance of each session, gender, age, height, and who was either an acquaintance or a friend of whom, were determined by a questionnaire which was handed out to the participants. The corresponding statistics on gender, age and height are given in table 14. All in all, the participants were distributed among five groups of two, seven groups of three, and five groups of four individuals. Table 15 provides an overview of group sizes and dyadic gender composition. Note the singular sample of female-only groups of two. In spite of the overall good ratio of 30:21 between male and female participants, their schedule, along with the schedule at which the infrared tracking system was available, led

		Gender		
Variable	Measure	Male	Female	All
Age	Mean	23.5	23.7	23.6
	Median	23	24	23
	StdDev	3.1	3.7	3.3
Height (cm)	Mean	181.0	170.1	176.6
	Median	182	170	178
	StdDev	6.2	5.5	8.0

Table 14.: Gender, age and height of the second experiment's participants.

	Arity			
Biological sex	2	3	4	$\sum$
male-male	2	5	10	17
male-female	2	12	8	22
female-female	1	4	4	9
$\sum$	5	21	22	48

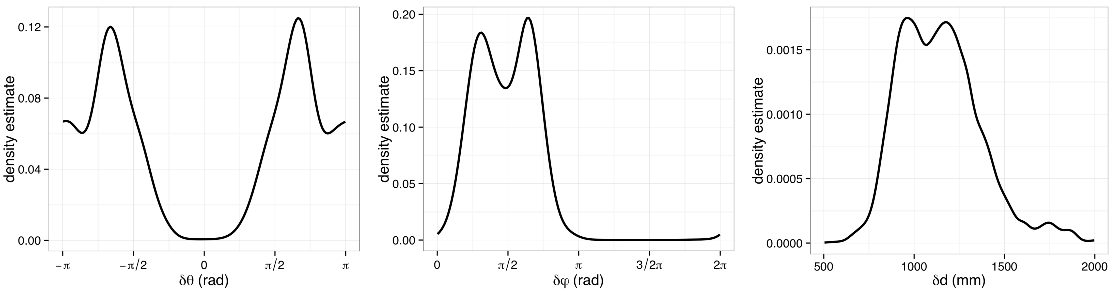
Table 15.: Dyads per group size and biological sex as in [81].

to this rather unfortunate sample size. One may also note that eight ouf of theoretically $5 \cdot \binom{2}{2} + 7 \cdot \binom{3}{2} + 5 \cdot \binom{4}{2} = 56$ dyads had to be removed due to marker failures.

Similar to the first experiment, each individual wore an infrared marker on their left or right shoulder, and the same infrared tracking system [8] was used to record the positions and orientations of the markers. Likewise, the recorded data were post-processed as described in sections 2.2.2 and 2.2.4, yielding position and orientation of each person at every time-frame. Finally, positions and orientations of the individuals led to the values of the introduced variables of interaction geometry, $\delta\theta$, $\delta\varphi$, and δd. The data of each session were annotated according to group size and gender of the interactants. According to the previous findings and discussions, it is expected that the distributions of $\delta\theta$, $\delta\varphi$ and/or δd are functions of either one or both of arity and gender. Density estimates of the variables of interaction geometry of the overall second dataset, as well as in regard of group size, gender, constant group size but varying gender, and constant gender but varying group size, are illustrated in figures 27, 28, 29, 30, and 31. All in all, the picture that shows is quite similar to the one from the first experiment for all variables. As expected, arity and gender both prove to be influential, which will be discussed in the following sections.

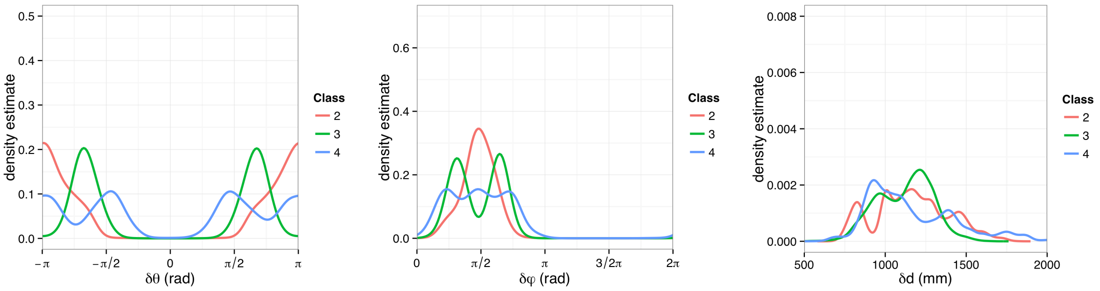
2.4.2.1 $\delta\theta$

The overall distribution of $\delta\theta$ is similar to the first experiment (figure 27a vs. 8a). Due to the restricted group size, the available space, and the assigned task, the movement dynamics were much less during the second experiment. This led to more clearly established FFSs, which also shows in the fact that almost no interactions were observed for $|\delta\theta| \leqslant 30°$. Among groups of two, three, or four individuals, the distributions of $\delta\theta$ exhibit clearly distinct peaks and variances (figure 28a). Interestingly enough, groups of two mostly engaged in full-frontal configuration, which they basically never did during the first experiment. It appears as if there simply was no good reason for other configurations because there were no significant spatial or (known) social factors present during the second experiment. More



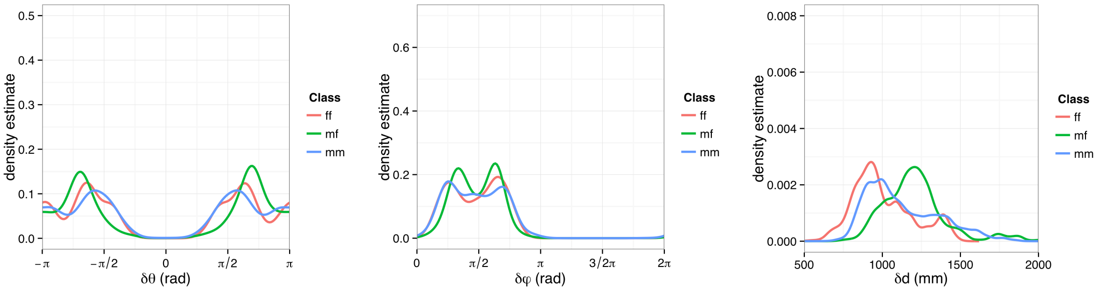
(a) (b) (c)

Figure 27.: Kernel density estimations of $\delta\theta$, $\delta\varphi$ and δd, using a Gaussian kernel and bandwidths of $10°$, $10°$ and $25mm$, respectively.



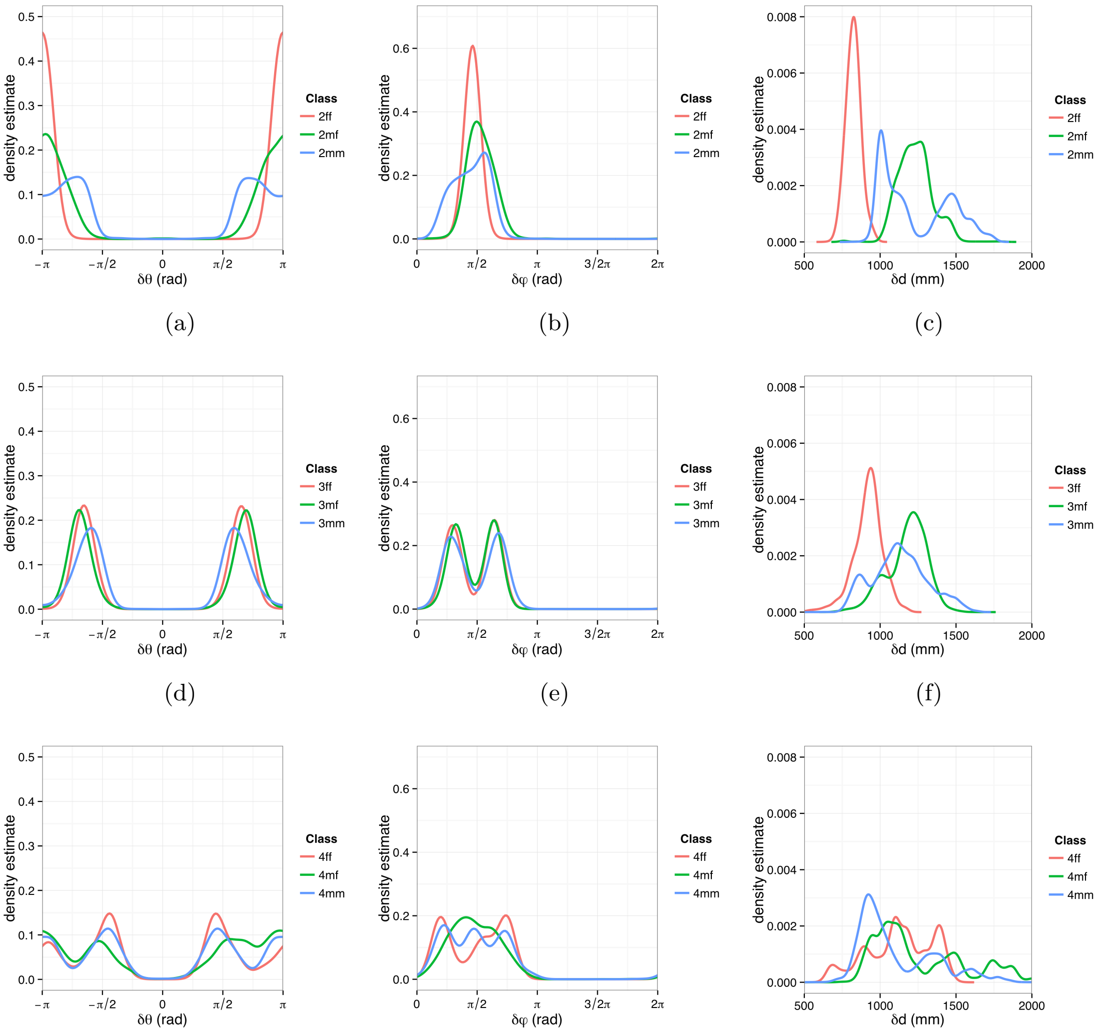
(a) (b) (c)

Figure 28.: Kernel density estimations of $\delta\theta$, $\delta\varphi$ and δd with respect to arity, using a Gaussian kernel and bandwidths of $10°$, $10°$ and $25mm$, respectively.



(a) (b) (c)

Figure 29.: Kernel density estimations of $\delta\theta$, $\delta\varphi$ and δd with respect to biological sex in dyads, using a Gaussian kernel and bandwidths of $10°$, $10°$ and $25mm$, respectively.

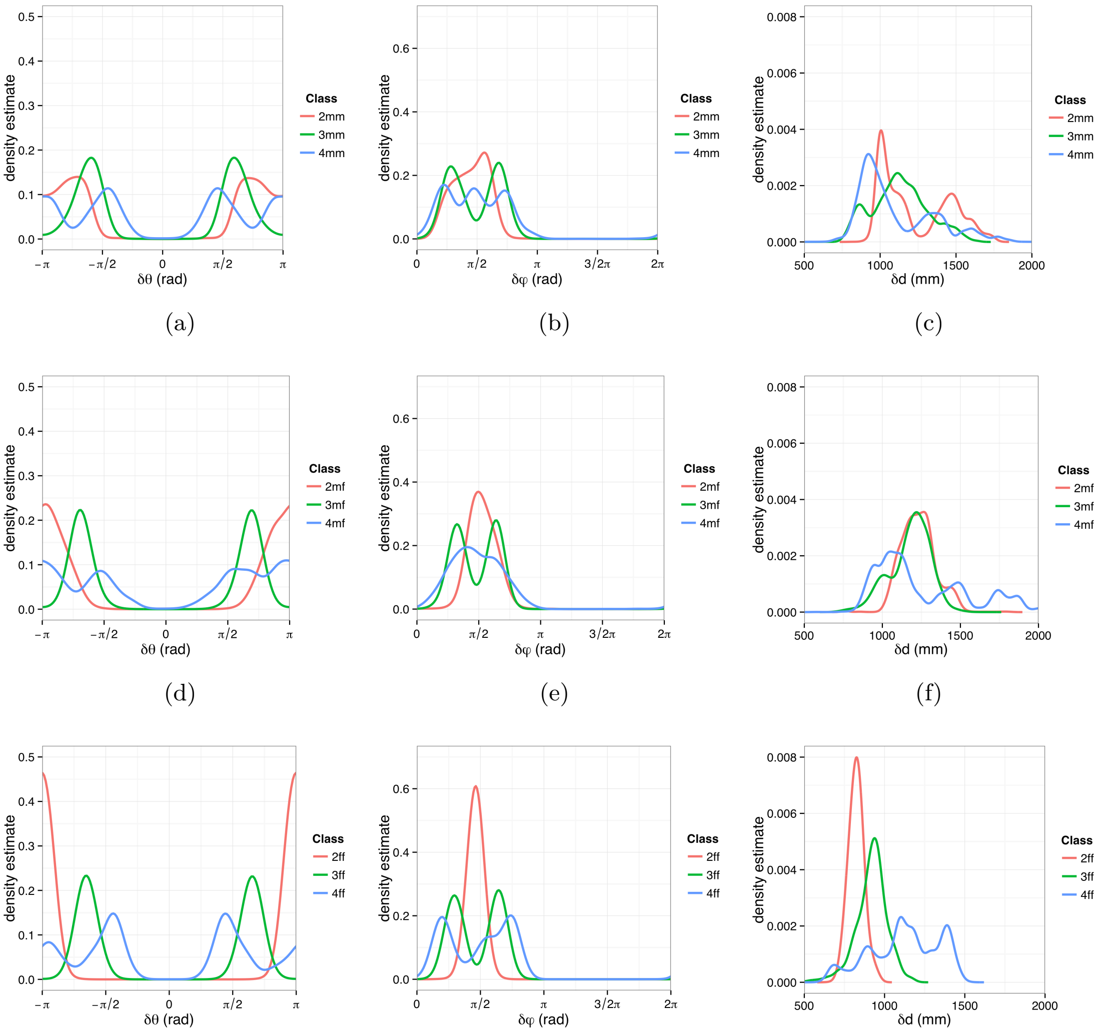


(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure 30.: Kernel density estimations of $\delta\theta$, $\delta\varphi$ and δd with respect to biological sex in dyads of groups of sizes two, three and four, using a Gaussian kernel and bandwidths of $10°$, $10°$ and $25\mathrm{mm}$, respectively.



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure 31.: Kernel density estimations of δθ, δφ and δd for same-sex dyads of groups of sizes two, three and four, using a Gaussian kernel and bandwidths of 10°, 10° and 25mm, respectively.

precisely, it seems plausible that, during the first experiment, groups of two more often chose L- or V-shaped F-formations in order to be able to monitor other groups and individuals, or display “openness” of their respective group so that others might join. The peaks for groups of three or four are close to the idealized values (refer to table 4, but the variance is notably less in both cases. In terms of $\delta\theta$, groups of four seem to adopt the most stable orientations. On the other hand, the probability of subgroups splitting off is fairly low for such small groups in any case.

On first sight, the distributions of $\delta\theta$ with respect to gender seem to be less distinct (figure 29a), yet still, for both peaks and variances, differences are apparent. For example, the peaks differ about ten to twenty degrees, which, taking into account the distance between a person’s left and right shoulder blades, will almost certainly (though unconsciously) be perceived by other interactants. Also, the most direct relative orientation towards each other was present in mixed dyads, while it was the least for male-male dyads.

Figures 30a, 30d and 30g illustrate the distributions of $\delta\theta$ within classes of equal arity. From these it follows that there a greater difference between dyads of varying gender in groups of two and four than in groups of three. The orientational behaviour in groups of two is especially interesting, as it further corroborates the notion that male dyads adopt less direct attitudes than female or mixed-sex dyads (figure 30a). Comparison with the graphs for other group sizes suggests that orientation with respect to gender should not be seen independent of arity like it used to be common practice in other related studies (see section 2.4.1.2). From a socio-psychological perspective, adopting less frontal configurations might be more important in groups of two in order to avoid subliminal aggressive behaviour [143]. Furthermore, orientation seems to be significantly more dynamic for mixed-sex than other dyads in groups of four. Lastly, comparison of mixed-sex with male-male and female-female dyads also reveals a tendency for clearly more frontal configurations for mixed-sex dyads, particularly so in groups of four (figure 30g).

2.4.2.2 $\delta\varphi$

According to figure 27b, overall no interactions took place in the rear of any person, which is as expected, again given the small group sizes and the experimental setting. In sum, most interactions occured at polar angles of 55° and 115°. Not unexpectedly, these peaks do not appear in the distribution of $\delta\varphi$ from the first experiment (figure 10) as the current distribution is biased towards groups of three and especially groups of two which, as discussed before, adopted full-frontal configuration in the second experiment, but basically never in the first.

$\delta\varphi$ varies significantly with arity (figure 28b). As expected, the peaks of $\delta\varphi$’s distribution for groups of two, three, or four, closely approximate the idealized values (refer to table 3), and variance follows group size. With respect to gender, mixed-sex dyads were located more directly towards each other at polar angles of about 60° and 114° as opposed to 46° and 121°. They also showed less dynamics (figure 29b).

Once group size is taken into account, the least variance shows for groups of two females, and the most for groups of two males (figure 30b). The shape of $\delta\varphi$'s distribution for the latter suggests that some men adopted more frontal positions than others. These findings should nonetheless be considered only with great care due to the small sample sizes for groups of two (see table 15).

Similar to $\delta\theta$, there is no significant difference between dyads of varying gender in groups of three (figure 30e). For groups of four, however, one notices a regular distribution of $\delta\varphi$ for male dyads, suggesting that the corresponding FFSs were constantly maintained (figure 30h). This is likewise the case for female dyads, yet it appears if some women stood closer together than others, probably as a result of their social relationship. In case of mixed-sex dyads in groups of four, the presence of local maxima at $80°$ and $110°$ indicate that one or more persons kept more distance to their opposite gender. The distribution of δd for groups of four (figure 30i) supports this picture because one would ideally expect three instead of four peaks. Obvious differences in proxemic behaviour can also be seen from the comparison of the distributions of $\delta\varphi$ for varying arity within classes of equal dyadic gender composition (figures 31b, 31e, 31h).

2.4.2.3 δd

The overall distribution of δd features two peaks at 962mm and 1175mm, regardless of group size or gender (figure 27c). As was previously discussed in the proceedings of the first experiment (section 2.2.5), basically no interactions take place at distances below 750mm or beyond 1500mm. The latter limit is obviously due to limits in group size and/or available space, whereas the former clearly follows "commonly agreeable" rules of proxemic behaviour which means that interactions at very close range are typically avoided.

According to figure 28c, δd's distribution has multiple peaks, namely at 825mm and, each about 10cm to 15cm apart, at 1010mm, 1162mm, 1263mm, and 1454mm. The first peak originates from a group of two turkish females who were also friends. While this peak exists in its own right, it should be regarded with care due to the nature of the social relationship in conjunction with small sample size (table 15). But even if this peak were canceled out, one can still see that the variation in distance is comparatively high for groups of two. In comparison, part of the peaks and variances in groups of three or four can be explained merely in terms of distributing N persons subject to e. g. circular formations. The latter distributions are shaped more clearly with peaks at 967mm and 1214mm (groups of three), or 929mm, 1086mm, and 1388mm (groups of four). This is interesting because the two peaks for groups of three differ in their magnitudes, meaning that more often than not one out of three individuals stood farther apart than the other two, which is not visible from the corresponding distributions of $\delta\theta$ and $\delta\varphi$ that both closely approximate idealized configurations (figures 28a, 28b). On the other hand, δd's distribution for groups of four follows basic expectations. What stands out for all group sizes is that, with a mean of 933.3mm and median of 967mm (canceling out the first peak in groups of two and regarding the first two peaks in groups of four as a single peak), the average distance in this experiment is greater by about 20cm than in the first experiment. This is certainly an effect of the additional available space since less persons crowded the room, but it also does not agree with related research where 50cm or more likely 70cm are considered as typical values [57, 79].

In the context of gender (figure 29c), the distributions of δd for mixed-sex dyads is especially different from those of male-male and female-female dyads. In mixed-sex dyads, the average distance between individuals of distinct sexes was significantly higher. The same is also apparent from the distributions for varying gender under constant group sizes (figures 30c, 30f, 30i), from which it follows that in groups of two or three mixed-sex dyads stood much farther apart than same-sex dyads. From the cases of arities two or three, the basic notion is that female dyads stand closer than male dyads, which in turn stand closer than mixed-sex dyads. All the same, there is non-negligible second peak at about 1500mm for male dyads in groups of two which contradicts generality. Also, this relation is not valid for groups of four, where male dyads stand closest, followed by mixed-sex and female dyads.

2.4.2.4 *Discussion*

Overall, the distributions of $\delta\theta$, $\delta\varphi$, and δd are similar to those from the first experiment, and the same holds once the dataset is split by group size. Compared to the first experiment, much more space was available for the participants, leading to an increased average interpersonal distance. Except for groups of two, it is supposed that the available space, the limited group size, and the experimental settings had no further influence on the concrete choice of the respective formations. Circular formations were however prevalent in both experiments.

Orientationwise, the least direct orientations were found for male dyads, followed by mixed-sex and then female dyads, regardless of group size. In this context, $\delta\theta$'s variance was noticeably less for mixed-sex than for other dyads. Recall that Jones had similarly reported that women adopted the most direct orientations towards others [157]. According to his results, male-male dyads showed the least direct orientations, which matches the present results, but then he found that mixed-sex dyads were less directly oriented than female-only dyads, as opposed to the current results. In a later study, Shuter took culture into account as well [301]. Jones' and Shuter's results line up for American citizens, whereas Shuter reports that in case of Germans, men and not women adopted the most direct orientations. Furthermore, Remland et al. also found that men in general were the least directly oriented, partially dependent however on age [266]. Contrary to Jones and Shuter, but in accordance with the current results, they suggest that mixed-dyads were the ones with the most direct orientations.

In regard of interpersonal distance, Dosey and Misels found that women would approach other women more closely than men, whereas men would would keep the same distances to men and women alike [78]. Hartnett et al. suggested that distance might also depend on a heterosexuality score, according to which men with a high score would allow women to approach them more closely [137]. Recall that Uzzell et al. stated that hitherto results of research on interpersonal distance were ambigious [330]. Nevertheless, related work seems to agree on the assumption that male-male dyads displayed the farthest interpersonal distances. Comparing this to the current results, it turns out that mixed-sex dyads in fact kept significantly more distance than same-sex dyads. The distributions of δd for female and male dyads turned out to be quite similar, but still suggest that female dyads stand closest. Also recall that territory grows with group size, even though not regularly [85]. This means that the space occupied and allocated by individuals may become less with increasing group size. The distribution of δd with respect to group size indeed suggests that this assumption might hold because in groups of four, the closest distances were found between adjacent members, provided that the small sample for female-female dyads in groups of two is canceled out.

At the bottom line, the current results should not be generalized in a way such that female dyads always stood closer than mixed than male dyads. The same applies to relative orientation. Nevertheless, the overall experimental setting can be arguably considered unconstrained in terms of space, and the behaviour of the participants during their casual conversations was reported to be spontaneous and without significant awareness of the experimental situation. From the results it is clear that arity has the most significant influence on all variables, but gender proved to be a non-neglibile factor as well, especially in mixed-sex as opposed to same-sex scenarios. The results furthermore suggest that there is a correlation between both variables, even though one might be tempted to regard them as independent. For example, one may note that territories occupied by females were comparatively smaller than those occupied by males (for illustration purposes, see figure 35). The variables' distributions with respect to arity, gender, or both, suggest that they can indeed be incorporated in an algorithmic model for the detection of social interaction, be it in the form of *a priori* profile parameters (gender), or as latent variables (arity).

2.4.3 *Evaluation*

The analysis of the newly acquired data has shown significant differences in the distributions of $\delta\theta$, $\delta\varphi$, and δd, for distinct genders and/or group sizes. Quality and quantity of those differences lead to the assumption that algorithmic models for social interaction will be capable of modeling and classifying such data accordingy, which will be further evaluated in this section. Again, the multimodality of the present distributions, along with the very good results from the evaluation in section 2.3.5, suggest the use of GMMs. For this, the data from the second experiment were partitioned according to gender, arity, or both, and the corresponding models were evaluated by 10-fold stratified cross-validation. Figure 32 illustrates the performance characteristics for each classification task. Models were computed for varying numbers of components, ranging from one to fifty. This was also done as additional means to double-check the basic decision towards GMMs with roughly ten components (see sections 2.3.4 and 2.3.5).

	Predicted					
Actual	mm	mf	ff	Precision	Recall	F_1-Score
mm	95612	40996	8702	64.8%	65.8%	65.3%
mf	34466	140646	14742	72.1%	74.1%	73.1%
ff	17386	13568	46568	66.5%	60.1%	63.1%

(a) Male-male, male-female, or female-female dyads (68.5% accuracy).

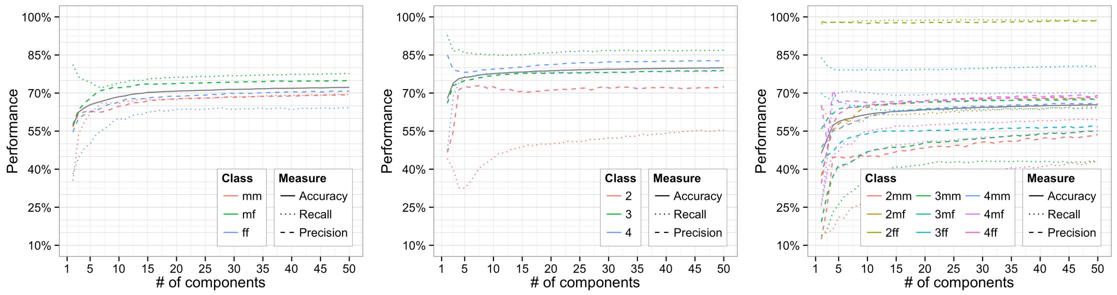
	Predicted					
Actual	2	3	4	Precision	Recall	F_1-Score
2	19107	12248	11651	72.5%	44.4%	55.1%
3	3210	160836	24782	76.8%	85.2%	80.8%
4	4032	36315	140505	79.4%	77.7%	78.5%

(b) Groups of two, three, or four (77.7% accuracy).

	Predicted											
Actual	2mm	2mf	2ff	3mm	3mf	3ff	4mm	4mf	4ff	Prec.	Rec.	F_1
2mm	4944	448	0	1289	5912	292	3777	1616	250	44.9%	26.7%	24.0%
2mf	197	10269	0	342	2739	13	1698	531	1225	64.4%	60.4%	62.3%
2ff	2	2	7356	0	7	83	2	12	0	97.4%	98.6%	98.0%
3mm	701	362	0	15625	16286	1695	5489	2944	1248	46.8%	35.2%	33.0%
3mf	2184	1487	9	4099	87708	7564	4136	1510	1993	65.1%	79.2%	71.5%
3ff	403	58	60	1884	5388	21731	3533	509	222	53.9%	64.3%	58.7%
4mm	1308	462	3	3690	6709	3504	57846	5552	3358	60.5%	70.2%	65.0%
4mf	1040	1751	128	3433	7426	3192	14095	28835	2250	66.1%	46.4%	54.5%
4ff	223	1118	0	3007	2646	2243	5107	2088	19838	65.3%	54.7%	59.5%

(c) Male-male, male-female, and female-female dyads in groups of two, three, and four (61.6% accuracy).

Table 16.: Confusion matrices after 10-fold stratified cross-validation of GMM-based classifiers on the second dataset.



(a) Male-male, male-female, and female-female dyads.

(b) Groups of two, three, and four.

(c) Male-male, male-female, and female-female dyads in groups of two, three, and four.

Figure 32.: Performance characteristics after 10-fold stratified cross-validation of GMMs with 10 components.

Figure 32a shows that gender could be classified in socially interacting dyads with an average accuracy of about 70%, for which the classifier chose the most likely model among models of male-male, male-female, and female-female dyads, solely based on the likelihood of the given observations under those models yet without prior knowledge of the true gender as input, thereby showing that gender-specific behavioural patterns are indeed characteristic. Precision and recall are comparatively higher for mixed-sex than for same-sex dyads. The fact that the classifier missed female-female dyads more often than others is compensated by its precision for that class. From the confusion matrix in table 16a it follows that male-male dyads were partially confused with male-female dyads, and vice versa, but rarely with female-female dyads. Female same-sex dyads, on the other hand, were slightly more often mistaken for male same-sex dyads than mixed-sex dyads. Considering the results for mixed-sex dyads, this suggests that the classifier is slightly biased towards interacting males, which is probably the case because the respective data show the most regular distribution, especially in comparison to females (see figure 29, further amplified by the much smaller class prior for female same-sex dyads.

Next, figure 32b features an average accuracy of about 80% for the discrimination of group size. In other words, samples of dyadic interaction are classified as belonging to a pair of persons within groups of two, three, or four. Precision and recall are about the same for the latter groups, but noticeably less for groups of two. The confusion matrix in table 16b shows that indeed a little more than 55% of the dyads in groups of two were mistaken for dyads in groups of three or four with an almost identical rate of failure. Vice versa, dyads in groups of three and four were classified as dyads in groups of four and three every now and then, but only seldom as groups of two. Groups of three have a significantly higher recall than the other two classes, whereas the classifier was most precise in case of actual groups of four. In accordance with section 2.4.2, the most discriminant properties of the variables of interaction geometry in the second experiment for groups of two versus other arities are their distinctly peaked distribution of $\delta\theta$ and $\delta\varphi$, indicating full-frontal

	Gender			Arity			Arity & Gender		
Measure	$\delta\theta$	$\delta\varphi$	δd	$\delta\theta$	$\delta\varphi$	δd	$\delta\theta$	$\delta\varphi$	δd
Mutual information	0.062	0.050	0.127	0.285	0.158	0.079	0.364	0.226	0.322
Uncertainty coefficient $U(class\|X)$	0.060	0.048	0.122	0.298	0.165	0.083	0.186	0.115	0.164
Symmetric uncertainty $(\frac{2\cdot I(X;\, class)}{H(X)+H(class)})$	0.071	0.042	0.105	0.339	0.140	0.068	0.272	0.138	0.194

Table 17.: Relevance of $\delta\theta$, $\delta\varphi$, or δd with respect to the class attributes.

and stable formations during the recordings. Nevertheless, similar to the prior discussion about the classification of female same-sex dyads, the sample size was much less than that for groups of three and four, and so is the class prior, thus effectively canceling out the aforementioned peaks.
The last of the three classification problems is concerned with the discrimination of varying gender dyads in groups of varying size. This task is therefore a nine-class classification problem. Figure 32c shows the results, where e. g. "3mf" represents the class of male-female dyads in groups of three. As the figure shows, the average accuracy is down to about 65%. This is nevertheless an acceptable result, regarding the results of the – so to speak – marginal classification problems. Overall, recall is relatively wide-spread among the numerous classes, whereas precision is generally closer to the average. It is not surprising that the corresponding precision and recall are "out of the roof" for female same-sex dyads in groups of two due its small sample size. In this nine-class problem, the distributions of $\delta\theta$, $\delta\varphi$ and δd are so characteristic for this class (see figure 30) that even the small class prior would not cancel out the effects on the overall model. Likewise, δd's distribution for dyads of varying gender in groups of three is especially characteristic for female-female and male-female dyads, but here the effect is canceled out due to the similarities for $\delta\theta$ and $\delta\varphi$. The confusion matrix in table 16c further illustrates that the classifier's performance was rather poor for classes "2mm", "3mm", and "3ff". Despite the slightly better performance for "4mm", one can detect a general tendency towards less performance for male-male dyads, regardless of arity. To the contrary, mixed-sex dyads are detected well, as similarly indicated by the results in table 16a. The overall results for female-female dyads are yet too ambigious with respect to varying arity to allow for generalization of the performance. Finally, it should be noted that the models show consistent performance in all three classification domains. Accuracy, precision and recall increase slightly with an increasing number of components. Especially the results of the nine-class problem suggest the use of more components. The objective, however, is not the further optimization of the classifier for this particular dataset, but achieving acceptable results for the general problem. In this regard, the demonstrated results corroborate the choice of about ten components for GMMs as algorithmic models for the detection of social interaction.

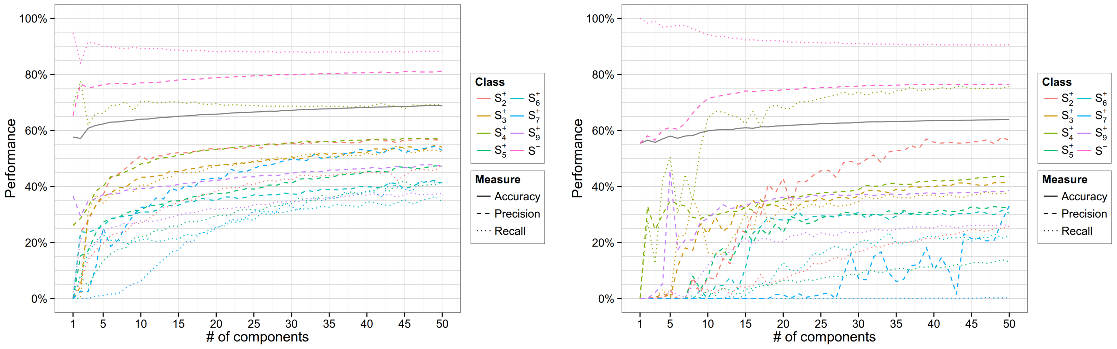
Similar to section 2.3.5.5, $\delta\theta$, $\delta\varphi$ and δd can be ranked in terms of their relevance for each of the three problem domains, as illustrated in table 17. According to the uncertainty coefficients, δd is predominant for the classification of dyads in regard of gender, followed by $\delta\theta$ and $\delta\varphi$ with roughly equivalent importance. This attributes to the related work and prior discussions which already assumed that differences in interpersonal distance were characteristic among both sexes. For the group size problem, on the other hand, the coefficients for $\delta\theta$, $\delta\varphi$, and δd are each about half of their predecessor, in that order. Recall that, in spite of the fact that all values were computed for unit-less variables and that the uncertainty coefficient is a supposedly normalized quantity, the underlying representation is still non-linear. Therefore it cannot be reasoned that one variable were twice as important as another. From the information-theoretical perspective one could however argue that on average twice as many *nats* are needed in order to convey the same information, given the class attribute. This perspective also allows for a comparison of the gender- and arity-problems in so far as that arity can be considered as more influential on interaction geometry than gender. For the nine-class problem, $\delta\theta$ and δd are equally ranked in terms of uncertainty, followed by $\delta\varphi$, even if the latter does not weigh significantly less. This can be losely interpreted as being accounted for by the "sum" of the separate informations according to gender and arity.

At this point, δd and $\delta\theta$ seem to be prevalent for the modeling of interaction geometry. The evaluation of the first experiment however showed δd and $\delta\varphi$ to be the most important variables, in that order (see table 13). Nonetheless, in both cases δd is the predominant factor, which also fits into the results from related work and the discussion from section 2.4.1. Especially in regard of the ranking of $\delta\theta$ and $\delta\varphi$, the overall sample sizes should be taken into account and therefore the results should not be generalized too much. Further experiments at much larger scale will help to determine which one has more utility, if at all. It is also likely that further experiments will show that the utility of the variables further depends on the social and/or physical context in which the respective transactions occur. For mobile agents, it is more difficult to measure $\delta\varphi$ than $\delta\theta$ or δd as it requires either precise knowledge of location on small spatio-temporal scales or equivalent, yet less accurate, means like the one presented in the remainder of this work (chapter 3).

2.4.3.1 *Reevaluation of the first dataset*

The previous section evaluated the performance of the newly acquired dataset for varying gender and group size. Due to the experimental design, these data lack a class equivalent to $S^{\ominus}$. Also, the problem domain was limited to groups of two, three, and four. Since the proposed models are generative, the lack of $S^{\ominus}$ could be compensated by drawing samples from the model computed for $S^{\ominus}$ on the previous dataset (see section 2.2.5). However, obtaining virtual data for $S^{\ominus}$ from the first dataset is not reliable in the gender-related context of the second dataset. After all, only two females participated in the first experiment. In regard of group size, though, the data from the first experiment can be re-evaluated instead. In addition to the availability of $S^{\ominus}$ and the comparatively greater sample size, this would also allow to analyze the task with respect to group sizes from two to nine (except for eight, refer to table 2 on page 28).

The first dataset was therefore split into classes $S_2^{\oplus}$, ..., $S_7^{\oplus}$, $S_9^{\oplus}$, $S^{\ominus}$, with class priors corresponding to the relative frequencies within the dataset. Table 18 features the confusion t



(a) Gaussian Mixture Model (GMM)

(b) Semi-Wrapped Gaussian Mixture Model (SW-GMM)

Figure 33.: Performance characteristics of GMMs- and SW-GMMs-based classifiers for a varying number of components after 10-fold stratified cross-validation on $S_2^{\oplus}$, ..., $S_7^{\oplus}$, $S_9^{\oplus}$, and $S^{\ominus}$.

matrix of this eight-class classification problem for a GMM-based classfier. Most notably, whereas overall accuracy is comparable to that of the three-class problem in the previous section, precision and recall are far from reasonable for all but $S^{\ominus}$. Comparing the results for $S^{\ominus}$ with those of the original evaluation (table 11 on page 73), one may notice a decrease of precision together with an increase of recall. This is unfortunate as it obviously implies an increase of false positives for the whole equivalence class of $S^{\oplus}$. Other than that, among $S_2^{\oplus}$, ..., $S_7^{\oplus}$, $S_9^{\oplus}$, the classifier least often predicted samples as $S_7^{\oplus}$, regardless of the actual class. This is explained by the small class prior of only $\sim 2.5\%$ in conjunction with the overly high variance of the samples for groups of seven (refer to the scatterplots in appendix B). In general, the higher the variance of the variables in one of $S_n^{\oplus}$, the more samples from the corresponding classes were predicted as $S^{\ominus}$.

A considerable fraction of the erroneous decisions occurred in favor of neighbouring classes. This is probably caused by the increasing variance of $\delta\theta$, $\delta\varphi$ and δd for groups of more than four or five individuals. In other words, this means that an increasing number of persons will position and orient themselves in more possible ways, especially in circumstances where the available space is limited, like it was the case during the first experiment. Eventually, this again leads to the question whether precision can be improved by increasing the number of Gaussians, or using SW-GMMs instead of GMMs, as some periodic properties might be attenuated only once the whole dataset is split in distinct $S_n^{\oplus}$. Once more, it should be noted that here the idea is not optimizing the classifier for this particular dataset, but finding out whether more Gaussians can possibly capture class-specific (in other words: arity-specific) effects. To evaluate this, GMM- as well as SW-GMM-based classifiers were computed for varying number of components, as illustrated in figure 33. Not unexpectedly, the results show only marginal improvements for an increased number of modes for both GMMs and SW-GMMs. It follows that there are no elementary differences between the variables' distributions for distinct group size in the "real world", or that

	Predicted										
Actual	$S_2^{\oplus}$	$S_3^{\oplus}$	$S_4^{\oplus}$	$S_5^{\oplus}$	$S_6^{\oplus}$	$S_7^{\oplus}$	$S_9^{\oplus}$	$S^{\ominus}$	Prec.	Rec.	F_1-Score
$S_2^{\oplus}$	4103	1491	5270	1289	70	11	1830	876	51.1%	27.5%	35.7%
$S_3^{\oplus}$	1318	11267	9821	2436	111	24	1552	1593	43.7%	40.1%	41.8%
$S_4^{\oplus}$	1065	3696	42566	4170	939	167	4954	2815	48.1%	70.5%	57.2%
$S_5^{\oplus}$	255	5648	10669	14294	2909	250	10552	19763	32.0%	22.2%	26.2%
$S_6^{\oplus}$	157	244	3586	5335	5136	525	8403	2594	31.2%	19.8%	24.2%
$S_7^{\oplus}$	3	315	2732	3575	904	1658	4911	11270	32.9%	6.5%	10.9%
$S_9^{\oplus}$	594	1485	8647	8790	3513	1079	41673	83331	38.7%	27.9%	32.5%
$S^{\ominus}$	538	1640	5275	4762	2865	1319	33788	407131	76.9%	89.0%	82.5%

Table 18.: Confusion matrix of $S_2^{\oplus}$, ..., $S_9^{\oplus}$, $S^{\ominus}$ after 10-fold stratified cross-validation of a GMM-based classifier (63.9% accuracy).

Measure	$\delta\theta$	$\delta\varphi$	δd
Mutual information	0.054	0.061	0.160
Uncertainty coefficient $U(\mathit{class}\|X)$	0.038	0.043	0.111
Symmetric uncertainty $(\frac{2\cdot I(X;\ \mathit{class})}{H(X)+H(\mathit{class})})$	0.040	0.046	0.115

Table 19.: Importance of $\delta\theta$, $\delta\varphi$, or δd with respect to $S_2^{\oplus}$, ..., $S_7^{\oplus}$, $S_9^{\oplus}$, and $S^{\ominus}$.

	Predicted				
Actual	$S_{combined}^{\oplus}$	$S^{\ominus}$	Precision	Recall	F_1-Score
$S_{combined}^{\oplus}$	245992	122242	83.1%	66.8%	74.0%
$S^{\ominus}$	50187	407131	76.9%	89.0%	82.5%

Table 20.: Confusion matrix of $S_{combined}^{\oplus}$ vs. $S^{\ominus}$ based on the results from table 18 (79.1% accuracy).

the sample size is insufficient to emphasize these differences. The former can certainly be neglected due to related work and prior discussion in sections 2.4.1 and 2.4.1.5.
Now that the data for $S^{\oplus}$ have been split into subclasses, the relevance ranking of the variables has changed (tables 19 and 13). After the split, the information gain of δd has significantly increased, whereas $\delta\varphi$ and δd can now be seen as conveying about the same amount of information. Recall that for the same problem in the previous section, $\delta\theta$ turned out to be the most dominant variable (table 17) while δd seemed irrelevant in comparison. These findings do however not contradict each other. Instead, they relay that the importance of relative distance is proportional to group size. Clearly, shoulder orientation and polar angle varied more in the second experiment when there were only small groups and no further constraints on spatio-orientational arrangements. During the first experiment, larger groups naturally occupied more space, and relative differences in shoulder orientation and polar angle tend to vanish with increasing group size.
So far, separate models per group size do not offer reasonable advantages. Aside from investigating separate $S^{\oplus}_{n}$, though, the question is whether – and if so, by how much – the overall task $S^{\oplus}$ vs. $S^{\ominus}$ may yet benefit from separate models per group size. As an idea, one could subsume the results for $S^{\oplus}_{n}$ from table 18 under a single virtual equivalence class $S^{\oplus}_{combined}$ (table 20), and then compare these numbers to those for $S^{\ominus}$. Doing so results in slightly better precision for $S^{\oplus}_{combined}$ in comparison to the results for $S^{\oplus}$ from the first experiment (refer to table 11), together with a sligh decay of recall. It may therefore seem as if favoring one model for $S^{\oplus}$ over combining multiple $S^{\oplus}_{n}$ is merely a matter of trading off recall for precision. A comparison of the results for $S^{\ominus}$ from tables 11 and 20 however reveals significantly less false positives for $S^{\oplus}_{combined}$ – as such a notable benefit, although one may argue that whether false positives outweigh false negatives is a matter of application-specific intent. Summing up, however, it was shown that the overall model can indeed be improved by incorporating arity.

2.4.3.2 *Posterior probability of group size*

It was mentioned that in addition to potential improvements of the model itself, the incorporation of latent variables such as group size could also help in the search for a function $p(n|x, \theta)$, yielding a probability distribution over the arity n, given a sample and a model of interaction geometry. A posteriori, such a function could e. g. provide auxiliary information for negotiations between two or more mobile agents about social situations.
From the previous section it is clear that computing one model per group size and classifying new samples according to these models yields poor precision and recall for the distinct classes. It was however shown that using this mechanism and combining the results into one virtual equivalence class $S^{\oplus}_{combined}$ yields an increase in precision, albeit at the cost of a few more false negatives, while the classifier's overall accuracy remains about equal in comparison to the first approach (refer to tables 11 and 20). Also, recall the assumption that the higher variance which goes along with both larger groups and $S^{\ominus}$ plays an important role in the results. Therefore the idea is to use a two-way procedure where first

	Predicted									
Actual	$S_2^{\oplus}$	$S_3^{\oplus}$	$S_4^{\oplus}$	$S_5^{\oplus}$	$S_6^{\oplus}$	$S_7^{\oplus}$	$S_9^{\oplus}$	Prec.	Rec.	F_1-Score
$S_2^{\oplus}$	4758	1630	4817	1245	91	35	2364	49.9%	31.8%	38.9%
$S_3^{\oplus}$	1578	11356	9520	3201	146	77	2244	44.4%	40.4%	42.3%
$S_4^{\oplus}$	1608	4234	41273	4793	1121	362	6981	50.5%	68.4%	58.1%
$S_5^{\oplus}$	412	5895	10637	19936	3013	556	23891	40.4%	31.0%	35.1%
$S_6^{\oplus}$	149	315	3584	4905	5374	744	10909	37.7%	20.7%	26.7%
$S_7^{\oplus}$	7	343	2908	3678	933	3752	13747	45.1%	14.8%	22.3%
$S_9^{\oplus}$	1022	1798	8919	11607	3578	2797	119391	66.5%	80.1%	72.7%

Table 21.: Confusion matrix of $S_2^{\oplus}$, ..., $S_9^{\oplus}$ after 10-fold stratified cross-validation of a GMM-based classifier (55.9% accuracy).

the data are classified as $S^{\oplus}$ or $S^{\ominus}$, and second those data which were predicted as $S^{\oplus}$ are further classified according to group size. Consequently, the one model for $S^{\ominus}$ will be disregarded during the second step, hence getting rid of a bit uncertainty.

Table 21 lists the results after cross-validation of a classifier built on $S_2^{\oplus}$, ..., $S_7^{\oplus}$, $S_9^{\oplus}$, but not $S^{\ominus}$. According to the results, the precision has increased for all groups of size greater than four, particularly so for groups of nine. Also, recall has improved for all classes, again even significantly for groups of nine. As a matter of fact, though, the classifier's overall performance is still far from satisfying. To overcome this issue one could of course further reduce the number of classes in an attempt to eliminate variance and corresponding uncertainty. In this regard, one could argue that groups of five or more persons are quite rare (see also section 2.4.3.3). And indeed, taking into account only the classes $S_2^{\oplus}$, $S_3^{\oplus}$, and $S_4^{\oplus}$ yields a notable increase in accuracy. Not unexpected, this is even more so than in comparison to just leaving out $S^{\ominus}$, where variance is high but the distribution of the samples differs more from $S_{2,3,4}^{\oplus}$ than $S_{5,6,7,9}^{\oplus}$ differs from $S_{2,3,4}^{\oplus}$. Altogether this shows that *a posteriori* information on group size is realizable, if only for smaller groups, which may yet be unsatisfactory. Table 21 however also reveals that part of the erroneous predictions happened in favor of *neighbouring* classes. One notable exception is $S_6^{\oplus}$ for which most samples were predicted as $S_9^{\oplus}$ instead. This is likewise the case for $S_5^{\oplus}$, yet less emphasized. Again, this is probably a consequence of the sample size, i. e. the relative short durations for which groups of five, six and seven were observed (see table 2). The distributions of $\delta\theta$, $\delta\varphi$ and δd for $S_2^{\oplus}$, on the other hand, have a lot in common with those for $S_4^{\oplus}$ (see appendix B). Thus, deciding between the latter two classes is often merely a matter of their class priors. The obvious question is whether this notion of frequent predictions towards adjacent classes can be exploited, for instance in terms of the expected value

$$\mathbb{E}\left[N|x,\Theta\right] = \sum_{n} n \cdot p(n|x,\Theta)\,. \tag{99}$$

Equation (99) obviously expresses what the classifier *momentarily* expects as group size n, given the present observation x and the models Θ that correspond to the respective arities. From this point of view, all results are equally "valuable", i. e. the classifier weights its decision according to the certainty or uncertainty of each model. Unfortunately, deciding for $S_i^\oplus$ when the real class is actually $S_{i+3}^\oplus$ is no different from deciding for $S_{i+2}^\oplus$, the latter being much closer to the truth. Decision theory, in comparison, is concerned with maximizing the outcome, or alternatively minimizing the loss, among two or more "actions" under the uncertainty of two or more future "states of nature", each of which may turn out to be true with a prior probability [223, 30, 235]. These probabilities may be derived from past observations or just as well be subjectively anticipated. For each pair of action and state, a value is then assigned that represents the *payoff* once that course of action were taken and that state effectively were to occur. As payoff may or may not be personally assessed in a possibly non-linear fashion, its value can furthermore be expressed in terms of its utility. For instance, someone might rate being given one million dollars for sure much higher than being given a 50% chance of winning three million dollars, although the expected values are in fact not far apart [235]. This approach could be transferred to the present problem as follows: Actions are given in terms of "choose $S_2^\oplus$", "choose $S_3^\oplus$" and so forth. States of nature then refer to the ground truth of the group size, and their prior corresponds directly to the posterior

$$p(n|x) \propto \sum_k p(n|x, \theta_k)p(k) \tag{100}$$

which was previously determined by the classifier. Let a_i denote the action "choose $S_i^\oplus$" and let s_n denote the state "ground-truth is $S_n^\oplus$". For each pair of action i and state n, the payoff is then given by $\nu_{ni} = 1/(1 + |n - i|)$. The latter introduces a penalty for increasing distance between chosen action and actual state, i. e. ν_{ni} represents an assessment of utility. The *payoff table* V is subsequently defined as follows:

		Decide for $S_j^\oplus$		
		$S_2^\oplus$	...	$S_9^\oplus$
Truth is $S_i^\oplus$	$S_2^\oplus$	$\frac{1}{1+\lvert 2-2\rvert}$	...	$\frac{1}{1+\lvert 2-9\rvert}$
	$S_5^\oplus$	$\vdots$	$\ddots$	$\vdots$
	$S_9^\oplus$	$\frac{1}{1+\lvert 9-2\rvert}$	...	$\frac{1}{1+\lvert 9-9\rvert}$

For every action, its expected outcome with respect to the state is thus determined by

$$\mathbb{E}_i\left[\nu_{Ni}|x, \theta\right] \propto \sum_{n \in N} \nu_{ni} \cdot p(n|x, \theta_n)p(n) \tag{101}$$

and hence the second step of the classification problem consists of finding

$$\mathit{argmax}_i \; \mathbb{E}_i\left[\nu_{Ni}|x, \theta\right] \; . \tag{102}$$

	Predicted									
Actual	$S_2^{\oplus}$	$S_3^{\oplus}$	$S_4^{\oplus}$	$S_5^{\oplus}$	$S_6^{\oplus}$	$S_7^{\oplus}$	$S_9^{\oplus}$	Prec.	Rec.	F_1-Score
$S_2^{\oplus}$	3964	1913	5418	1721	194	30	1700	55.9%	26.5%	37.3%
$S_3^{\oplus}$	1128	11376	9996	3696	172	81	1673	45.1%	40.5%	42.4%
$S_4^{\oplus}$	930	3845	43486	5065	1211	351	5484	50.8%	72.0%	59.2%
$S_5^{\oplus}$	172	5570	10704	22860	2977	673	21384	39.0%	35.5%	34.5%
$S_6^{\oplus}$	88	335	3584	6363	6048	843	8719	37.2%	23.3%	25.8%
$S_7^{\oplus}$	1	272	2754	4198	1087	4662	12394	46.6%	18.4%	21.6%
$S_9^{\oplus}$	810	1920	9686	14772	4593	3357	113974	68.9%	76.4%	72.7%

Table 22.: Confusion matrix of $S_2^{\oplus}$, ..., $S_9^{\oplus}$ after 10-fold stratified cross-validation of a two-step GMM-based classifier and maximum expected payoff (56.0% accuracy).

The performance results of this two-way procedure are shown in table 22. From these it follows that precision got better merely for $S_2^{\oplus}$, while recall increased for all classes except $S_9^{\oplus}$. Notably less samples from other classes were predicted as $S_2^{\oplus}$, yet more instances from $S_2^{\oplus}$ where mistaken for $S_4^{\oplus}$. The performance of $S_5^{\oplus}$ has considerably improved in so far as that much more instances were correctly predicted and also much less were mixed up with $S_9^{\oplus}$. For $S_3^{\oplus}$, the "distance" to erroneously predicted samples decreased, and also much less $S_3^{\oplus}$ were mistaken for $S_9^{\oplus}$. Nonetheless, too many $S_3^{\oplus}$ are still misclassified as either $S_4^{\oplus}$ or $S_5^{\oplus}$. Last, just like before $S_6^{\oplus}$ are often seen as $S_5^{\oplus}$, but again notably less as $S_9^{\oplus}$. All in all, the two-step procedure has helped to reduce uncertainty, in particular between instances of $S_9^{\oplus}$ and the remaining classes, and has furthermore brought any misclassified instances closer to the ground truth. Unfortunately, though, the overall performance increase through maximum expected payoff is insignificant in comparison to the standard *maximum posterior* approach. It should also be noted that much of the classifier's accuracy is due to the comparatively large number of samples from $S_9^{\oplus}$.

Surely the results are better than making random decisions between the seven classes, and the predictions could possibly still be used as additional input for mobile agents that negotiate about social situations. The standalone performance is however not enough for reliable a posteriori information on group size. The presented procedure, which borrowed its idea from decision theory, is considered a first approach to solving this problem. Even though only marginal improvements were observed, they suggest further research in this area. In spite of the fact that simple adaption of the payoff ν_{ni} in form of a simple exponential decay did not yield significant results, adapting values and distribution of payoff as well as further research on the class priors seems promising. Section 2.4.3.3 discusses prior probability distributions of group size in more detail.

As identification of group size is primarily a problem of at least two agents, it would also make sense to include the (uncertain) knowledge of other agents into the process. This augmentation of decision theory towards maximization of utility among multiple agents can then be seen as a problem of game theory. As samples are always measured with

respect to dyads, the first logical step into this direction could consist of incorporating the partner's measurements of the variables which should always be available, or else could be estimated as discussed in section 2.2.5.4. At the same time this would yield the possibility to include other class priors or even other specific models, for instance those which were learned by other agents during their lifespan.
What certainly has to be taken into account with respect to the present work is that each of the models for $S_2^{\oplus}$, ..., $S_7^{\oplus}$, $S_9^{\oplus}$ features rather distinct peaks which are specific for the respective group size, whereas there often is a natural overlap in the underlying data. Due to these overlaps, it is expected that a posteriori estimation of group size will in any way not improve by much for this particular dataset. Generally speaking, the more dynamic or, to the contrary, the more limited a situation is, e. g. in terms of physical constraints, the more likely such overlaps will be attenuated. The latter does however not contradict the general case of modeling interaction geometry. As far as *dynamics* in social situations are concerned, an alternate approach will be presented in chapter 5.

2.4.3.3 *On class priors for group size*

Section 2.3 discussed that, among other things, one advantage of generative classifiers is their incorporation of class priors. The posterior probability of a class, given an observation and a set of class-dependent model parameters, is determined by its likelihood, weighted by the prior probability of the class. Under Bayes' Rule, classifiers then choose the class with the maximum posterior. From a *frequentist perspective*, these class priors can be won by analyzing the relative frequencies of each class in a dataset, as opposed to *Bayesian inference* which models uncertainty of variables in the form of probability distributions in their own right [218]. For the evaluations in sections 2.3.5 and 2.4.3, the class priors were computed based on the number of occurrences of each class in the respective datasets because the overall number of observed subjects and groups is rather small and hence these priors resemble the objective truth during the experiments. The distributions of the class priors should nevertheless be generalized in order to arrivate at a preferably universal model. As a matter of fact this turns out to be a very difficult problem. The class priors for males and females, for instance, could be chosen according to demographic studies, from which it follows that e. g. a German is about 1.05 times more likely a woman than a man [3]. World-wide, it is however 1.01 times more likely the other way around. The census furthermore shows a significant change of this ratio along with increasing age [4]. Coming back to social interaction detection, it is also quite clear that such a class prior strongly correlates with the actual social situation and its environment. However, following the discussions in section 2.4.1 this comes to no surprise.
Social networking has attempted to overcome these matters by means of *random graphs*, based on the seminal works by Erdős and Rényi [87, 88]. A random graph $G_{N,M}$ of N nodes and M edges is constructed via drawing one out of $\binom{\binom{N}{2}}{M}$ equiprobable graphs. Another, recently more widely adopted formulation is that of a random graph G_N, whose vertices represent individuals and whose edges represent domain-specific social links, such as the

fact that adjacent individuals participate in the same social situation, and where each of the $\binom{N}{2}$ pairs of nodes is connected with probability $0 \leqslant p \leqslant 1$, i. e. all edges have the same probability and all are drawn independently of each other [40]. Thus the average number of edges is expected to be $\binom{N}{2} \cdot p$.

Since each draw is equivalent to a Bernoulli trial, the probability of a vertex v with degree d follows a binomial distribution:

$$p(deg(v) = d|N, p) = \binom{N-1}{d} p^d (1-p)^{N-1-d} \tag{103}$$

It can be shown that, if $N \cdot p \to \lambda$ for constant λ, $N \to \infty$, and $p \to 0$, this distribution approximates a Poisson distribution:

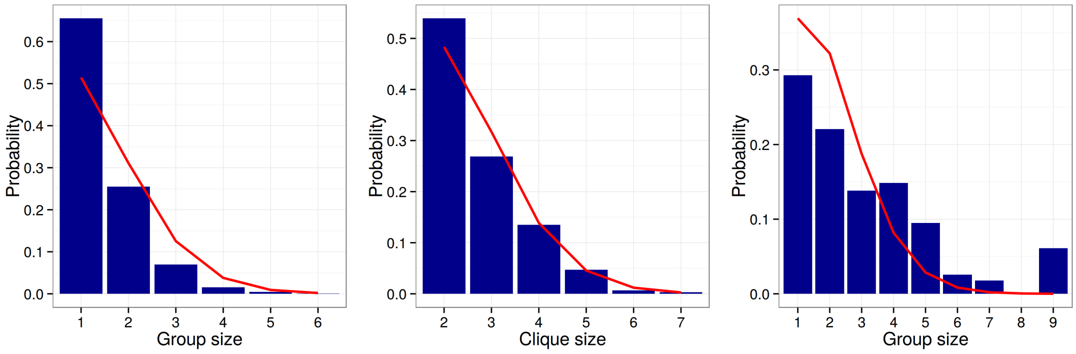
$$p(deg(v) = d|\lambda) = \frac{\lambda^d}{d!} e^{-\lambda} \tag{104}$$

This kind of distribution of the vertex degree is in fact often found in related works [152, 217, 222, 219]. Erdős, Rényi, and Bollobás have shown several important properties for these graphs, such as the expected number and size of cliques as a function of $N \cdot p$, or consequently the probability $p(k) = (1-p^k)^{N-k} p^{\binom{k}{2}}$ that k vertices span a clique in G_N, i. e. a complete subgraph of k vertices.

It is tempting to use equation 104 for modeling the prior probability of group sizes in social interaction, for which groups of one are interpreted as individuals which are not part of any group. From this point of view, groups of zero hence do not exist. For this, a zero-truncated Poisson distribution compensates for the missing zero by scaling the remainder of the distribution as follows:

$$p(deg(v) = d|\lambda) = \frac{1}{1-e^{-\lambda}} \frac{\lambda^d}{d!} e^{-\lambda} \tag{105}$$

Poisson distributions have been used for modeling group size [152, 79] (see figure 34). They fit both distributions [152] and [79] well (see figures 34a and 34b), which can e. g. be verified via Pearson's χ^2-test for goodness of fit. To the contrary, Moussaïd et al. [217] reported that Poissons were a suitable match for only one out of two populations in their research, which they assume to be a consequence of the distinct environments in which the observations were made. In the context of random graphs, emphasis should be put on sufficiently high number of vertices, in turn relating to sample size. While sample sizes were arguably high in the studies from James and Dunbar, they were considerably less (in terms of groups) in case of Moussaiïd [152, 79, 217]. Also, note that Dunbar did not report any statistics with respect to groups of one. Similarly, groups of one did not occur during the second experiment of the present work, albeit as a result of the experimental settings. The distribution of groups in the first experiment is in principle similar to both [152] and [79], yet in spite of the presence of groups of one, corresponding to individuals outside of or transitioning between groups), the distribution is rather heavy-tailed and therefore not a good match for a Poisson (see figure 34c). Apart from Poisson distributions, James also fitted negative binomial distributions and found that the latter were a significantly



(a) Group sizes according to [152]. (b) Clique sizes according to [79]. (c) Group sizes from the first experiment.

Figure 34.: Distribution of the group sizes plus fitted zero-truncated Poisson distributions (red).

better match for about 94% of the 18 populations which he had observed [152], whereas Poissons would match only about 61% at the same significance level ($p < 0.05$). A negative binomial distribution is of the form

$$p(k|r) = \binom{k+r-1}{k}(1-p)^r p^k \tag{106}$$

and can be understood as the probability of observing $p(k|r)$ groups of size k until encountering r groups of different sizes. James argues that the mean of a Poisson is constant in spite of the fact that environments and settings might change between observation periods [152]. He therefore assumes that a Poisson would “fit those distributions from social situations where the relationships governing the combinations of individuals were relatively stable” [152]. A negative binomial, on the other hand, were are supposedly better match for “diverse empirical distributions” since it could subsume a “family of different Poisson distributions” [152]. The relation between the negative binomial and the Poisson distribution can be shown analogously to equation (104) by considering the limit $r \to \infty$ for constant $\lambda = r\frac{p}{1-p}$.

At the bottom line, none of the presented approaches is yields a panacea for the issue of modeling class priors. Related models are based on nothing but random graphs for which the probability of each edge has the same probability and is drawn independently from the others. In a recent review, Newman et al. reason that random graphs “turn out to have severe shortcomings as models of such real-world phenomena” [222]. From the studies of numerous social (and other) networks, it follows that in many cases vertex degree is unlikely to follow a Poisson distribution, so that one has to mind the possibility that important properties of the corresponding networks are being ignored once such distributions were used [222]. To the contrary, social networks derived from real-word data are often found to obey *heavy-tailed distributions*. Heavy-tailed distributions exhibit a “heavier” tail than the exponential distribution, hence their name. These properties express themselves in great skewness and/or kurtosis. Among the family of heavy-tailed distributions, the *fat-tailed distributions* typically follow the power law, stating that one variable changes as the power of another. Fat-tailed distributions are widely found in social networking theory [219]. It is in fact an important insight that actual distributions from studies are mostly far from random [219]. Coming back to the experiments that were presented as part of this work, it is without doubt transparent that the distributions of group size and gender were strongly influenced by the experimental settings, more precisely the fact that the gender of dyads was not chosen randomly, and neither random are the observations of groups of nine individuals. Moreover, most random graphs fall short of the ground truth since in general they consider only undirected edges. Clearly, it can make a difference how probable it is to have an edge from A to B as opposed to an edge from B to A. For example, recall the finding that pairs of persons judged their mutual relationships differently in terms of being good friends or merely acquaintanced [143]. For these reasons, Newman et al. postulate the generalization of modeling towards non-poisson distributions and therefore investigate directed graphs, bi-partite graphs, and random graphs with arbitrary distributions of vertex-degree [222].

Yet another phenomenon which is commonly found in social networks is that of *triadic closures*, describing that strong ties from A to B and A to C imply at least a weak tie from B to C [118]. Social network analysis makes use of this in terms of the *clustering coefficient* in order to quantify the degree of clustering of a particular graph [341, 222]. The clustering coefficient, for example, plays an important role for the *small world model* by Watts and Strogatz [341]. In comparison to modeling random graphs, the small world model is based on the notion that the distance between vertices may actually relate to e. g. geographic or social distance, and that in such cases the probability of being connected tends to be higher for many scenarios. The model thus minimizes the average path length between vertices, whereas it maximizes the cluster coefficient [341, 219].

Summa summarum, it follows that class priors may be generalized under certain conditions, one of which is invariance of environment and/or context over the course of the observations. A lot of research has been done on random graphs, and many (social) networks have been successfully modeled this way. Among others, negative binomial, Poisson and fat-tailed distributions are predominant in the field, but both, distributions and their parameters, have to be chosen with great care and are usually application-specific. Newman goes as far as saying that the field still had to be considered as being "in its infancy" [219]. In regard of the present work, the modeling of class priors is therefore an open question albeit an essential one. It is already clear that, just like profile parameters, the distributions of gender, group size, or even the choice of F-formation are significantly influenced by multiple factors such as the environment, the purpose of a social transaction, or personal parameters. For this work, the relative frequencies of the classes have been used to model prior probabilities since they reflect the ground-truth during the respective experiments. In the attempt of finding a preferably universal model, more experiments will have to be conducted in order to determine one distribution that minimizes the error in relation to the actual distributions. It is highly likely that no such single distribution can be found and hence that distributions of class priors have to be chosen depending on other data that are known to involve (mobile) agents, and that distributions more often than not must be chosen according to specific applications.

2.4.4 *Discussion*

The last part sought to investigate options for improving the model for social interaction geometry through integration of profile and/or other parameters. Corresponding questions were how *a priori* knowledge could be integrated, and if so, to what extent would it yield improvements for the underlying problem of determining interaction as evidence for social situations. Lastly, it was investigated whether the model could be improved in a way such that information like group size can be inferred *a posteriori*.

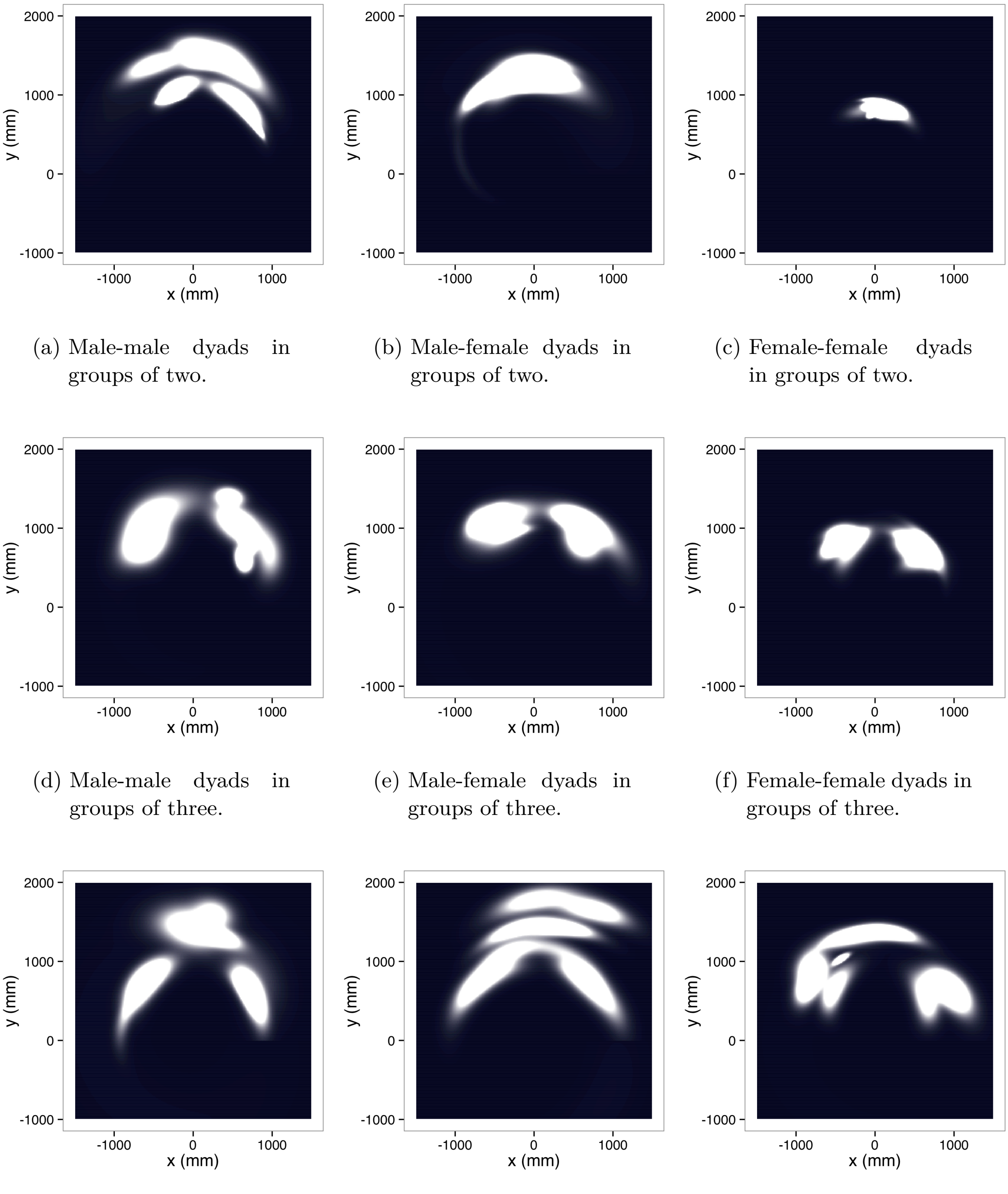
According to the related work in socio-psychological research over the last decades, it was determined that a multitude of latent and non-latent variables influence proxemic behaviour at various degrees. Among those variables that have sincere and noticeable effects on proxemics are for instance culture, gender and age. Aside from such more or less personal profile parameters, physical and other environmental constraints naturally have their share in altering proxemic behaviour. These parameters are often difficult or perhaps impossible to compute or understand in their entirety. Nevertheless, a particularly intuitive and rather expressive non-personal variable is given in terms of the number of interactants in a social situation. A small number of related papers have investigated group size from different perspectives and under varying circumstances. Their findings and possible consequences for the distribution of prior probability of arity were discussed in section 2.4.3.3.

Since gender and arity were considered as the most suitable variables because they can easily be quantified, and especially because they are (mostly) unambigious, a second series of experiments was conducted in order to determine if, in general, these two parameters convey enough information to have considerable influence on the algorithmic model for social interaction geometry. Indeed, in accordance with related work, the variables $\delta\theta$, $\delta\varphi$, and δd feature characteristic conditional distributions with respect to gender and/or arity. It is certainly justified to consider related work as ambigious in this regard. Likewise, no overly simplifying hypotheses should be concluded from the observed distributions of the variables, such as the presumption that women would generally interact at closer ranges than men. Nevertheless, the distributions are distinct enough to safely assume that, even though not explainable by simple heuristics, they still vary under gender and/or arity alike. This is corroborated by the information gain of each of the variables with respect to the classes, from which it e. g. follows that distance is predominant for distinct behaviour of men and women (in the present data). More generalizing statements would clearly require much bigger datasets, and due to the high diversity of human behaviour, as well as the influence of further latent variables, it is also doubtful that individual (marginal) variables, such as shoulder orientation or relative distance alone, will categorically obey any kind of generalization. On the other hand, the integration of multiple variables of interaction geometry does in fact show interesting properties which also align with the common notion in related work, for instance that women claim smaller territories than men or how groups of different sizes are likely to adapt certain formations. For the second dataset, this is illustrated by figure 35. Likewise, the data from the first dataset were analysed with respect to varying arity in order to find further confirmation, and also because naturally occurring groups of up to nine people were observed in the original experiment while cardinality was a controlled parameter in the second experiment. Note that whereas figure 36 illustrates the distinct zones of interaction for groups of up to nine individuals, it also confirms that variance and overlap are proportional to arity, which has been identified as one major negative impact on classifier performance for a posteriori information about group size.

In a corresponding series of evaluations of the second dataset, the dataset was partitioned according to gender and/or arity, and models were computed based on GMMs. More precisely, the performance characteristics were determined for classifiers built on top of models for male-male, male-female, and female-female dyads, and/or groups of two, three, or four participants (see table 16). The results show that gender and arity can indeed be inferred, with acceptable accuracy in case of gender and noticeably better accuracy in case of group size. Combining gender and arity, and hence computing separate models for each of $\{mm, mf, ff\} \times \{2, \dots, 7, 9\}$ resulted in comparatively poor performance, particularly for male-male dyads in groups of two or three. It was presumed that this is a consequence of both the variables' distributions as well as the class priors, suggesting that subsequent experiments ought to be conducted for further clarification. At the bottom line, though, the results prove that both gender (more precisely biological sex) and arity have in fact significant influence on proxemic behaviour, and thus should be respected by algorithmic models of social interaction geometry.

The first dataset was subsequently reevaluated according to group size. This dataset is about twice the size of the second dataset and features groups of up to nine individuals. Also, the second series of experiments was designed such that all subjects were continuously engaged in mutual interaction ($S^{\oplus}$), so that this reevaluation allowed for embedding arity in the larger context of $S^{\oplus}$ vs. $S^{\ominus}$. Due to the fact that for some group sizes the number of samples is still small, which is not unexpected since larger groups naturally occur less often (refer to section 2.4.1.5), and because some observations lie close to the periodic limits of $\delta\theta$ or $\delta\varphi$, classifiers were now based on either GMMs or SW-GMMs for $S^{\oplus}_{2}$, ..., $S^{\oplus}_{7}$, $S^{\oplus}_{9}$, and $S^{\ominus}$. Arguably, the overall accuracy of the classifiers was acceptable at about 64%, but precision and recall were insufficient. SW-GMMs would not perform better than their non-periodic counterparts.

Further analysis of the confusion matrix (see table 18) revealed that erroneous predictions often occurred in favor of adjacent classes and that the higher variance of the variables for bigger group sizes ($n \geqslant 5$) played an important role in the decision process. Following the discussions in section 2.2.5, the variance is partly due to spatial constraints during the first experiment, but naturally also due to increasing "degrees of freedom" along with increasing group size (figure 36). The video footage, for example, features an occasion where a group of three stood very close to a group of two, effectively reflecting the interaction geometry of a group of five, and very likely a consequence of the limited space. In any way, such corner cases do exist and should not be neglected. Proxemic behaviour in groups of



(a) Male-male dyads in groups of two.

(b) Male-female dyads in groups of two.

(c) Female-female dyads in groups of two.

(d) Male-male dyads in groups of three.

(e) Male-female dyads in groups of three.

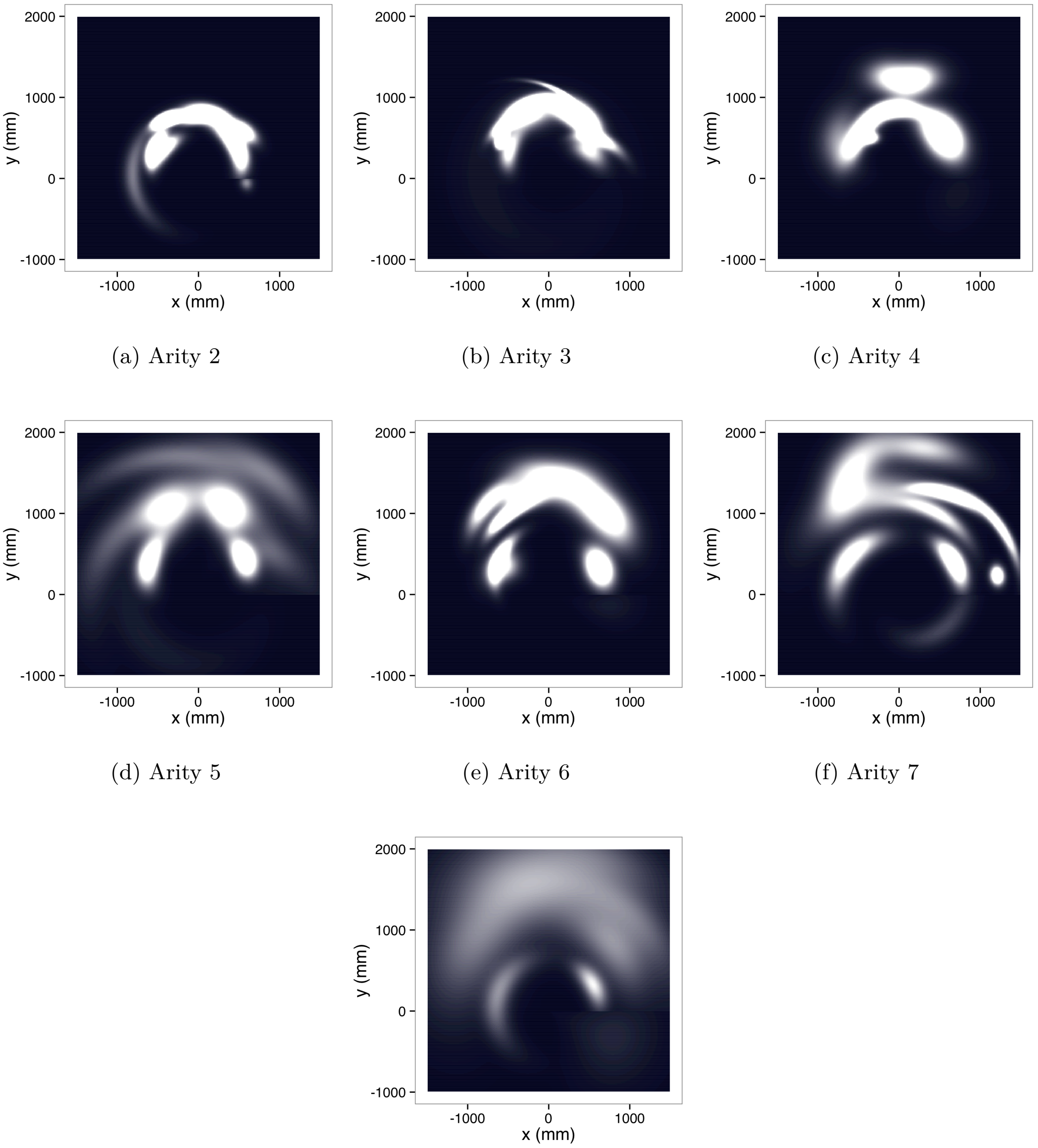
(f) Female-female dyads in groups of three.

(g) Male-male dyads in groups of four.

(h) Male-female dyads in groups of four.

(i) Female-female dyads in groups of four.

Figure 35.: Orthographic projection of the intensity of social interaction according to group size and gender, based on models corresponding to the second dataset.



(a) Arity 2

(b) Arity 3

(c) Arity 4

(d) Arity 5

(e) Arity 6

(f) Arity 7

(g) Arity 9

Figure 36.: Orthographic projection of the intensity of social interaction according to group size, based on models corresponding to the first dataset.

two, on the other hand, is presumably subject to greater influence of variables like gender or mutual social relationship than in bigger groups. Nevertheless, it is expected that particularly the characteristics of groups of five or more members would be attenuated once more data were sampled. This is also reasonable in the common sense that more degrees of freedom require more data [34, 218], when seen in terms of possible formations and variations of proxemic behaviour instead of only variables of interaction geometry. As the set of $S_2^\oplus$, ..., $S_7^\oplus$, $S_9^\oplus$ is equivalent to $S^\oplus$, the results were then subsumed under a single virtual class $S_{combined}^\oplus$ which, in the larger context of presence vs. lack of social interaction, led to an increase in precision for $S_{combined}^\oplus$ in comparison to the original evaluation (tables 18 and 11), albeit at the cost of recall. Whether applications would trade off recall for precision or rather stick with the original approach is certainly domain-specific. In the end, the increase in precision adds to the assumption that incorporating parameters such as arity, e. g. by means of multiple specific models, tends to improve the overall approach. At this point, one notable finding is that relative distance becomes increasingly important once group size is taken into account, not only when seen from an information-theoretical perspective but also explainable by the understanding that relative orientation and polar angle become less important at scale with increasing group size. This is also why the results of evaluating the second dataset do not contradict those from reevaluating the first dataset. According to the former, $\delta\theta$ and $\delta\varphi$ conveyed the most information whereas this is not the case for the latter. As the second series of experiments was restricted to groups of two, three, or four, the comparatively higher ranking of $\delta\theta$ and $\delta\varphi$ is clear, just like the insight that this effect vanishes the more groups grow in size. Another, yet very important, result is certainly given by the finding that groups of two, three, or four individuals behaved very much alike in the first experiment and the second series of experiments although they were totally unrelated. This strongly attributes to the hypothesis on the generalizability of algorithmic models of social interaction geometry.

The next research question of this section was concerned with a posteriori information about group size. So far, the classifier was designed with the main goal of predicting presence or lack of social interaction from samples of *dyadic* transactions. Even though the dataset was partitioned according to group size, and corresponding models were learned for the reevaluation with respect to group size (and possible improvements of the principle $S^\oplus$ vs. $S^\ominus$ problem), group size is a priori unknown. Among other things, mobile agents might benefit from a posteriori knowledge about group size, for instance when negotiating on the whole set of persons who are the supposed members of a particular social situation. The classifier performed poorly with respect to discriminating between all of $S_2^\oplus$, ..., $S_7^\oplus$, $S_9^\oplus$, and $S^\ominus$. To the contrary, satisfying results were shown for the general classification of data according to $S^\oplus$ respective $S^\ominus$, with about 80% accuracy and arguably acceptable precision and recall of both classes. Therefore the idea was using a two-fold procedure where the first step is only concerned with discriminating between $S^\oplus$ and $S^\ominus$ (or $S_{combined}^\oplus$ and $S^\ominus$ for that matter), and the second step further processes the certainty or uncertainty of the first classifier about group size for an improved prediction thereof. Computing the expected value for group size would not suffice because at least some of the distributions of $\delta\theta$, $\delta\varphi$ and δd have too much overlap among groups of different arity. In particular, those classes that feature *high* intra-class sample variance will likely yield GMMs with highly variant components and therefore mostly evenly distributed probability density, as opposed to classes with *low* intra-class sample variance. In the latter case, the probability density is more likely to have several characteristic modes, but inbetween those modes the probability density would be comparatively low. As a consequence, the likelihood of the models for bigger group sizes will most certainly not decline below some threshold, and therefore the expected value will suffer from the weights of those classes with higher intra-class sample variance.

For these reasons, an alternative approach was presented, based on the idea of maximizing the "payoff", like for example postulated in decision theory. Payoff was quantified as a function of the distance between chosen and possible group sizes. These values were weighted with the prior probabilities with which groups of different arities occur, and finally the group size with the maximum value for expected payoff was chosen. This way it was possible to reduce the misclassification rate between neighbouring classes which consequently showed in slightly increased precision and recall, even though the overall accuracy would remain about equal to the results based on maximum posterior selection (tables 21 and 22). The results support the reasoning that a posteriori estimation of group size can be realized. In spite of the fact that the classifier is far from being usable in terms of precision, it is still way better than random, and its predictions, or at least the underlying models' likelihoods, might nonetheless serve as auxiliary evidence in the negotations of mobile agents. The presented approach should be regarded as a first step in this direction, and future work should e. g. continue with a more detailed analysis and modeling of the class priors, or with finding ways to reduce intra-class sample variance respective ways to attenuate the specific characteristics per group cardinality. Generally speaking, for this particular approach towards posterior estimation of group size, bigger groups will likely continue to pose a problem. Instead one might think of alternatives like considering only those observations from e. g. the three mutually closest dyads.

There is no doubt about the influence of profile parameters such as gender or latent variables like group size, and that the determination of social interaction can indeed benefit from incorporating such parameters. The remaining question is thus concerned with the modalities of incorporating additional knowledge into algorithmic models for social interaction geometry. Biological sex, for example, is a personal profile parameter. As such, it is easily available a priori knowledge for all individuals in question of interaction. The fact that this variable is dichotomous suggests that one could learn distinct models for men and women, or with respect to the present work, distinct models for male-male, male-female, and female-female proxemic behaviour in dyads. The appropriate models could then be selected prior to predicting whether two corresponding individuals interact. Abstracting over this idea leads to a decision tree where e. g. gender-specific decisions at the root conclude which models are chosen at the next layer. This concept is easily generalized for further parameters. On the other hand, the disadvantages of this approach are quite obvious. First, agents would likely have to store numerous models in order to cover each and every possible parameter setting. For the basic case of gender in dyads, this leads to six models ($\{\mathsf{mm}, \mathsf{mf}, \mathsf{ff}\} \times \{\mathsf{S}^{\oplus}, \mathsf{S}^{\ominus}\}$). Incorporating further variables yields exponential growth, subject to their respective discrete domains. This issue can be somewhat compensated by means of techniques such as decision tree pruning [212]. It is also put into perspective by considering the fact that due to the choice of GMMs for interaction geometry, the actual number of parameters for each model is very low and linearly dependent on the selected number of components. Other than that, one could also think of homographies from *specific* variables of interaction geometry to *generalized* variables of interaction geometry. Consider for example a fictional finding from which it followed that women would always locate themselves at closer ranges, and else their zones of interaction, e. g. due to mutual relationship, would differ from a general model only in terms of scale. While this is arguably far-fetched, one can still imagine that translation, scaling, and possibly rotation of the observations of $\delta\theta$, $\delta\varphi$, and δd might lead from a specific to a general model. If something like that were possible, it would considerably reduce the necessary number of models respective parameters. Second, in spite of pruning or the low number of model parameters, the inclusion of additional variables and their (discrete) domains naturally imply more degrees of freedom, therefore also requiring exponentially more training data. Third, there is always a risk that personal profile parameters are (intentionally or unintentionally) configured with the wrong values, thereby introducing systematic errors and bias. Profile parameters may also be fuzzy or ambigious. Likely examples are gender, age, or particularly "culture" (refer to section 2.4.1). Fuzziness is an interesting property and might perhaps be exploited. Recall that the inclusion of profile parameters is primarily supposed to aid in the discrimination of presence or lack of social interaction. It is furthermore evident that proxemic behaviour, aside from the assumption that there is a "greatest common divisor" among humans, has its corner cases and that peoples' behaviour sometimes just does not comply with what is expected under given circumstances (or profile parameter settings, for that matter). The overall classification of $S^{\oplus}$ and $S^{\ominus}$ might hence benefit from looking "past the edge", so to speak, e. g. by means of a weighted average or likewise a majority vote between models at adjacent nodes in a tree.

It is obvious that further research is necessary to achieve possible and sustainable modalities for the incorporation of additional parameters in algorithmic models for social interaction geometry. Parameters must be chosen with great care and with respect to the corresponding application domain, for which only those parameters should be considered that yield a significant information gain. Also, the introduction of specific additional variables may potentially be a consequence of relying on certain heuristics, which may or may not enhance the model. However, heuristics are generally subject to some sort of interpretation of a problem domain. Consequently, they may constrain the understanding of the general domain according to the specific interpretation of the problem domain.

3 POSITION AND ORIENTATION OF INDIVIDUALS

3.1 INTRODUCTION AND RELATED WORK

So far, the construction and evaluation of the proposed model were based on data gathered from surface-mounted motion capturing devices which tracked infrared markers worn by the subjects who participated in the experiments. For the model to be applicable in real world scenarios, it is however substantial to find means of measuring interaction geometry that do not depend on any external infrastructure, such as e. g. computer vision equipment, GPS, or any active or passive fixated sensors. Instead, mobile agents should be employed with *self-sufficient* techniques. Nevertheless, this does not necessarily imply that other sensors or techniques should not be taken into account when they are available and could actually help to reduce uncertainty, for instance in regard of location estimates. As discussed in chapter 1, present-day mobile phones – in particular *smart phones* – feature numerous physical and logical sensors, for instance accelerometers, magnetometers, gyroscopes, GPS receivers, Bluetooth, wireless networking, barometers, thermometers, near field communication devices, and potentially also ultrasound senders and receivers [288]. Smart phones are capable of continuously sampling, interpreting, and providing sensor measurements, which they do in a very unobtrusive manner, and prove to be a good source of information when it comes to the detection of social interaction in terms of interaction geometry. Modern Application Programming Interfaces (APIs) abstract from raw sensor output to single variables or rather complete models of orientation angles in degrees, acceleration in meters per second squared, pedometers, and so forth. In spite of advanced APIs, however, accurate position and orientation estimates still pose a series of complex problems, for example as introduced by sensor drift, bias, precision, sample rate, quantization, calibration, alignment, non-orthogonality, non-linearity as well as numerous more systematic and random error sources [188]. In the context of interaction geometry, orientation and location of a mobile device furthermore have to be related to orientation and location of the user.

This chapter will start with an overview of past and present techniques for estimating mobile device attitude and position, upon which the wearing habits of mobile phone users will be discussed. Finally, new approaches for relating mobile phone orientation to the user's body with respect to a global reference frame, as well as for measuring distance, to some extent also including relative position (in terms of $\delta\varphi$), will be presented and evaluated.

3.1.1 *Orientation*

In an early study, Mizell [213] used accelerometer measurements to estimate a vector parallel to the direction of gravitational force, and furthermore determined the dynamic components of acceleration, regardless of device attitude. For this, the accelerations along three orthogonal axes were sampled over the course of a few seconds and the samples were averaged. Except for situations where the device would be in free fall or subject to enormous acceleration (aside from gravity), the average vector v is parallel to gravitational force. Since the horizontal plane of the accelerometer, and hence the device, is perpendicular to v, the device's attitude is therefore determined up to roll and pitch, yet with one degree of freedom left, i. e. the angle about the yaw axis. Picking one of the samples at random and subtracting the estimated v yields the dynamic part d of the measured force. The vertical component p of d with respect to the device's attitude can then easily be determined by projecting d onto v. It follows that the horizontal component is given by $h = d - p$. Note that the direction of h is ambigious, so that further processing would be restricted to its magnitude $|h|$. Kunze et al. have augmented this idea in order to predict complete device orientation solely based on accelerometer readings [179]. For this, they employed simple heuristics according to which the acceleration along a pedestrian's walking direction is supposed to be highest next to gravitational force. Instead of averaging over the samples, they use a sliding window technique to determine the gravity vector v whenever total variance is close to zero and magnitude approaches $9.81\ ms^{-2}$. The horizontal plane is defined through its perpendicularity to v. Samples from the accelerometers are then projected onto the horizontal plane, and the walking direction is determined as the first principal component, i. e. the first eigenvector of the covariance matrix of the projected samples. This leaves questions about the relative orientation between device and body as well as the absolute orientation of the device with respect to some East-North-Up (ENU) global reference frame. Based on the previous work, Henpraserttae et al. [142] determined a transformation from sensor signals at arbitrary orientations into a global reference frame, and report fundamental improvements on classification performance for activity recognition tasks.

Other techniques determine global device orientation from a fusion of acceleration and magnetic field measurements. For this, acceleration as well as inclination of the earth's magnetic field are typically measured along three orthogonal axes. These measurements already yield information about two principal axes of the local coordinate system, since accelerometer measurements are always subject to gravity so that a vector pointing to the earth's center is easily determined, as is a vector pointing at magnetic north derived from magnetic dip. The third axis is then determined as the cross-product of these two vectors, and finally either one of the first two vectors is recomputed as the cross-product of the other two for orthogonolization, and all vectors are normalized. The former is also known as the TRIAD algorithm [36, 300]. Nevertheless, this method is clearly susceptible to disturbances in the magnetic field, additional acceleration apart from gravity, and clearly suffers from singularities at locations close to the magnetic poles. Such issues can be somewhat compensated through the additional incorporation of gyroscopes. Gyroscopes measure rate of rotation in degrees per second or similar units. Integration over time consequently leads to angles of rotation. Recall that rotation is not commutative, i. e. the order of rotation matters. A typical rotation sequence is the yaw/pitch/roll sequence, commonly found in applications for aircraft- or spacecraft attitude estimation. It can be shown, however, that the order of rotation does not matter for infinitesimal angles [188], from which it follows that very high sample rates are mandatory in order to keep measured angular rates at a minimum. Gyroscopes are not influenced by magnetic disturbances or acceleration, which is why they are often fusioned with accelerometer- and magnetometer-based systems. They are nevertheless susceptible to sensor drift and other systematic or random errors. In spite of the fact that the fusion of gyroscopes and other sensors for improved attitude determination had been known for a long time [170], Barthold et al. were among the first to exploit the fusion of *low-cost* gyroscopes with accelerometers and magnetometers on consumer mobile phones [27]. In order to cope with drift they determined the average drift of the sensor at different orientations over a series of measurements and subsequently used the results for corrections during the actual integration process. It should be noted that even though they used low-cost Micro-Electro-Mechanical System (MEMS) sensors, operated at low samples rates of only 8 Hz for the accelerometer and magnetometer, as well as 100 Hz for the gyroscope (note the difference), their estimations were predicted within 6% of the ground truth.

Further improvements include the use of complementary filters, i. e. combined low-pass filters to cope with short-term influences on accelerometer and magnetometer readings together with high-pass filters which are supposed to compensate for long-term drift of the gyroscopes, as well as linear or extended Kalman Filters (KFs), and finally also the use of quaternion algebra to avoid singularities such as gimbal lock [174, 17, 94, 276, 58, 43, 344, 323].

3.1.2 *Position*

Indoor localization techniques are commonly based on a subset of *time of flight* of signals, various kinds of *fingerprinting* and *dead reckoning*. Bahl and Padmanabhan were among the first to introduce a radio-frequency based system called RADAR with which a user's location could be estimated "within a few meters of his/her actual location" [20]. This system consisted of three base stations at three distinct locations on an office floor, operating at 2.4 GHz. Samples of signal strength and signal-to-noise ratio were collected throughout their scenario. These samples were then used for comparing actual measurements against the sampled data for estimations of the user's location. The authors report a median resolution of two to three meters, from which they conclude that their and equivalent systems are likely suited for applications at course room-level granularity. This finding is further corroborated by [7] in [95], according to whom RF-based methods cannot achieve accuracies below one meter due to their high frequencies and the lack of appropriate high precision timers and measuring equipment in consumer hardware.

In similar fashion, Bluetooth-based indoor localization systems were investigated in a number of studies [46, 24, 197, 52]. Bluetooth devices transmit at different power levels and hence cover different ranges, the common range being considered as up to ten meters, although practically the effective range is often much less due to distortions and reflections of the signal. For enhanced resolution of signal-strength as measured by the mobile devices, Bandara et al. [24] therefore proposed access points which would attenuate the signal, thus achieving accuracy within two meters in up to 72% of their test cases. On a general side-note, Bluetooth devices are assigned the roles of either master or slaves, where one master can handle a maximum of seven slaves. The latter is an important fact in regard of social interaction detection, as it would impose an undeniable limitation on cardinality, despite the fact that groups of seven or more individuals are rather unlikely (figure 26a). Also, devices are required to engage in a bonding through one of various pairing mechanisms, of which most require some sort of user interaction during the process.

Next, WiFi-based methods also have a long history in indoor localization [54]. Based on WiFi, positions are either determined via trilaterion of received signal-strength or by means of fingerprinting. The latter is similar to RADAR [20] because it requires that the signal strengths (and possibly additional features) of all access points which can be received at a set of discrete locations are recorded in advance to estimating a user's position. It follows that fingerprinting can only provide position estimates with high average error since the position of the best matching access point is chosen as the current position, according to distant metrics between the actual and the previously recorded measurements [92]. Time of flight, on the other hand, can theoretically provide better estimates of the user's location via trilaterion, but the proposed logarithmic models which relate signal-strength to distance are rather simplistic [54] as they do not consider any disturbances, reflections, or signal multihop. To some extent, these issues can be compensated with more or less sophisticated filtering techniques such as KFs or Particle Filters (PFs) [92, 50]. Evennou and Marx [92], for instance, report average measurement errors of 2.56 meters for KFs as well as 1.86 meters for PFs along a trajectory inside a building with four WiFi access points. Much like WiFi fingerprinting, *magnetic field fingerprinting* has been exploited for the purpose of indoor navigation based on the notion that the earth's magnetic field is characteristically disturbed by structure in buildings, installed equipment, power lines, water pipes, and so forth [313, 55, 104]. Chung et al. recorded the deviation between measured and actual heading along the corridors inside a laboratory building [55]. For every known location, three-axis magnetometer measurements were determined at four different orientations around the yaw-axis, namely 0°, 90°, 180°, and 270°. These measurements were used to build a map so that later the location of a device could be determined by finding the one location with the closest fingerprint according to Root Mean Squared (RMS) distance, for which they report a mean prediction error of about three meters and standard deviation of about four meters. As a side-effect, since the fingerprints had been recorded at four different headings each, the device's orientation about the yaw-axis could be determined with a mean angle difference of about 4° and standard deviation of about 5°. Similar errors were reported in a larger setting of two buildings with multiple floors, both buildings connected via pathways. Galván-Tejada et al. [104] improved this method by using a Fourier transform of the signal, where the analysis of the transformed signal in terms of its energy signature relaxed the information gathering process in so far as to help get rid of the need to sample measurements in several orientations about the yaw-axis. As a matter of fact, though, they only evaluated how the system performed with respect to recognizing rooms and therefore coarse granularity. Somewhat related to the prior techniques, Prigge and How [251] used multiple low-frequency magnetic field beacons throughout a building [251], while Pirkl and Lukowicz propose *magnetic resonant coupling*, employed as a system of surface-mounted transmitter coils in conjunction with mobile receivers [239, 240]. The system is characterized by an oscillating magnetic field, thus effectively reducing disturbances through even large metallic objects. Depending on the number of transmitter coils, they report quite accurate measurements, ranging from 44 cm accuracy with 33 cm standard deviation for one coil to 4 cm accuracy and only 6 cm standard deviation with four coils. Nevertheless, aside from the mandatory infrastructure this method also depends on precisely synchronized time which may be hard to achieve in autonomous mobile scenarios.

Moreover, a number of acoustic- and optical-based techniques have been published. Azizyan et al. presented a method for labeling *logical locations* such as “Starbucks” or “McDonalds” [18]. Similar to WiFi or magnetic fingerprinting, this method is based on the assumption that each location has its own characteristic photo-acoustic fingerprint. Their SurroundSense framework combines optical, acoustical and motion sensors for estimations of the user’s location, for which they report an accuracy of up to 87%. A very similar technique has later been labeled as *acoustic background spectrum* by [317]. Likewise, Constandache et al. [59] have developed a system that is able to compute routes between any pair of persons, provided that the walking trails of different individuals (among them naturally also the respective pair) have been learnt together with where and when they usually encounter. In addition to these data, they do however also require a fixed audio beacon for global reference.

Different from techniques that rely on external infrastructure, Peng et al. have presented a highly accurate solution for measuring the distance between two devices using only the standard microphones and speakers of consumer-level mobile phones [229]. It is remarkable that their method is neither subject to errors due to uncertainty in time synchronization, nor any misalignment between timestamp and actual signal emission, nor the time that goes by between receiving a signal and recognizing it as such due to delays caused by hardware and/or software. For this, both devices record incoming audio signals. The first device sends and subsequently receives its own signal. The same signal is also received by the second device, which in turn sends another signal in response, also eventually received by both devices. The algorithm then works as follows: Let t_0, t_1, t_2, t_3 denote the times at which the first device receives its own signal (t_0) as well as the subsequent signal from the second device (t_3), whereas the second device receives the first signal (t_1) followed by its own (t_2). Both devices measure the amount of time (in number of samples) between receiving their own and the respective other signal, independently of each other. Let these be denoted as

$$\delta t_{\mathit{first}} = t_3 - t_0 \quad \text{and} \quad \delta t_{\mathit{second}} = t_2 - t_1 \ . \tag{107}$$

The times of flight of the signal from the first to the second device and vice versa therefore are

$$\mathit{tof}_{\mathit{first,second}} = t_1 - \hat{t_0} \quad \text{and} \quad \mathit{tof}_{\mathit{second,first}} = t_3 - \hat{t_2} \ , \tag{108}$$

where $\hat{t_0} = t_0 - \mathit{delay}_{\mathit{first}}$ and $\hat{t_2} = t_2 - \mathit{delay}_{\mathit{second}}$ account for the tiny delays between sending and receiving ones own signal. These delays are system specific and can be determined a priori. It follows that

$$\begin{aligned} \delta t_{\mathit{first}} - \delta t_{\mathit{second}} &= t_3 - t_0 - (t_2 - t_1) \\ &= (t_3 - t_2) + (t_1 - t_0) \\ &= \mathit{tof}_{\mathit{second,first}} + \mathit{tof}_{\mathit{first,second}} + \mathit{delay}_{\mathit{first}} + \mathit{delay}_{\mathit{second}} \ . \end{aligned} \tag{109}$$

So the difference between $\delta t_{\mathit{first}}$ and $\delta t_{\mathit{second}}$ amounts to the doubled distance plus the delays between the two devices. It is trivial to derive the actual distance in units like meters or feet in relation to sample rate and speed of sound. Note that assuming e. g. a typical sampling rate of 44.1 KHz and speed of sound $346 \ ms^{-1}$ at $25°C$, the minimal distance that could be measured would be roughly $0.8 \ cm$. Also note that it is not necessary to express distance in specific units. Knowledge of the exact speed of sound, which varies e. g. with temperature, is therefore not essential. Based on [229], Filonenko et al. [95] discuss the feasibility of a system for trilaterion. They refer to Borriello et al. [41], according to whom it is possible to emit and receive sound signals at 21 KHz from standard consumer hardware such as mobile phone speakers, which is slightly above the range perceived by humans. Trilaterion would however require at least three devices, and if the *absolute* position of another device were needed, precise positions would have to be available for these devices. If a trilaterion system were designed in analogy to [229], it would furthermore be necessary to synchronize these devices. Existing ultrasonic trilaterion systems would therefore most often use a centralized approach which in turn requires infrastructure, e. g. in the form of a “dense grid of sensors on the ceiling” [95].

Finally, Dead Reckoning (DR) is a well-studied technique which estimates the current position by constantly updating the last known position according to speed and direction of travel. It is based on Newton’s laws of motion, according to which bodies maintain their state of motion unless external force is applied, and if that happens, changes happen proportional *to* as well as *in* the direction of the acting force [188]. In other words, keeping track of directions and magnitudes of the forces which act upon a body allows for extrapolation of its position. Simplistic systems therefore keep track of the heading and number of steps that a person has taken in order to estimate their current location, be it indoors or outdoors. People perform an average of 8265 steps per day under light to moderate activity, or 11603 under structured vigorous activity [343]. Personal step size has been found to be remarkably constant [158]. In 1997, Judd presented his first dead reckoning module consisting of three-axis accelerometers and magnetometers [158]. The magnetometer was used to determine the heading along which a person would be walking, whereas the accelerometer allowed for counting the number of steps as well as computing the horizontal plane, the latter of which is deemed especially important in areas where the vertical magnetic dip would exceed horizontal magnetic dip [158]. Step size and rotational offset of the device relative to the body could be manually entered, but also be automatically determined by means of a KF, provided that additional GPS signals and hence "ground truth" were available [158, 156]. The latter technique furthermore allows for estimation of the local magnetic variation, i. e. the angular difference between magnetic and true north. In a related work, Randell et al. [261] compared different configurations of sensors and sensor placements, and were able to perform step-based DR with a cumulative error of about four meters and standard deviation of about two meters over a walking distance of 126 meters. They also mention the additional use of gyroscopes for attitude estimation which, for example, were later also integrated in the NavShoe system for pedestrian tracking [98]. Link et al. did not use gyroscopes but improved location estimates via sequence alignment algorithms from bioinformatics [192], whereas Jin et al. fusioned DR estimates from two independent sets of sensors (Android-based mobile phones) as a constrained optimization problem for which they report error reductions of up to ~ 74% [155].

Instead of keeping track of the user's current *position*, Blanke and Schiele [37] predicted *transitions* between known locations in an office with reasonable accuracy, regardless of placement and orientation of the mobile device. They used a two-step procedure where first body motion was used for rough estimates of the device's orientation during one-second intervals, and second the differential principal component of three-dimensional rotation samples from the gyroscope would indicate the predominant heading vector when projected onto the ground plane (as determined by the first step). This method is based on the assumption that the main axis of rotation is determined by rotations and movements of the limbs, regardless of whether the device is carried in the hand or in the pocket (figure 37). In total they achieved very good results even though they assumed constant speed during motion and consequently the system failed when e. g. a person turned without moving forward at the same time. Back in the context of plain DR, Steinhoff and Schiele later found DR performance to be only slightly worse for arbitrary placement and orientation of a mobile phone in comparison to a "well calibrated, dorsally fixated sensor" [312]. Li et al. further employed particle filters and evaluated their system with more than fifty subjects over an accumulated distance of more than forty kilometers, for which they report mean errors between 1.5 and 2 meters dependent on whether the phone was carried in the hand or in a trousers pocket [190].

The NavShoe system [98] finally went beyond the principle of step counting in favour of Inertial Navigation (IN), which is but closely related to DR. IN is concerned with keeping track of position with respect to a chosen inertial reference frame, for which Inertial Navigation Systems (INSs) constantly measure acceleration and rotation rates of a rigid body (the tracked device) on three orthogonal axes. INSs have advanced from strap-down systems, consisting of a plate which was strapped down to e. g. an aicraft fuselage and

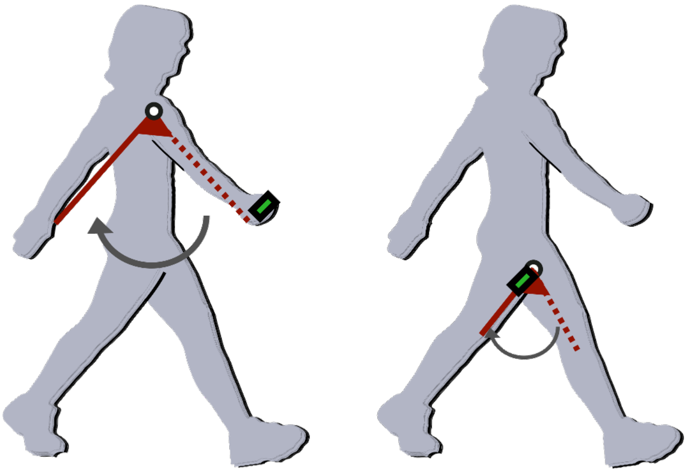


Figure 37.: Dominant rotational component when walking (figure taken from [37]).

kept level by means of mechanical gyroscopes, to miniature integrated systems of several MEMS dies [170]. Note that the measurements are performed at very small time intervals, for extremly accurate knowledge of the body's orientation within the reference frame is mandatory. In the context of of acceleration measurements, it is furthermore important to differentiate between *specific* and *total* force, since the former acts only relative to the reference frame whereas the latter is subject to gravity [43, 344, 323]. Due to the fact that INSs estimate position in terms of double integration of acceleration over time, it is clear that linear errors in acceleration end up as cubic errors in position [349, 188]. Therefore even the slightest error in estimated orientation, say 1°, and consequently the acceleration measurements projected into the estimated rotational frame, after only thirty seconds yield a positional error of $\sin 1^\circ \cdot 9.81 \text{ ms}^{-2} \cdot 30\text{s}^2 \approx 5 \text{ m}$. In addition to the application of Extended Kalman Filters (EKFs), NavShoe managed to reduce the cubic error to linear scale by introducing a simple heuristic, namely feeding "zero-velocity updates as pseudo-measurements into the EKF" during stance phases, based on the notion that walking consists of phases of stationary stance and moving stride, both of which last about half a second in turns [98]. Nevertheless, at present, INSs can still be considered unfit for high-performance indoor location, at least with respect to desired sub-meter or hopefully sub-decimeter accuracy, even if heuristics such as the one employed in the NavShoe or otherwise sophisticated sensor fusion are employed [98, 92, 350, 347].

3.1.3 *Orientation and location relative to the body*

Algorithms either require location and orientation to be known or must be invariant to both [178]. It is evident that orientation and location of the mobile device alone are not sufficient for exhaustive use in social interaction geometry. Instead, the device must be related to the user's body in order to determine mutual distance (δd), angle between

between shoulder lines ($\delta\theta$), and perhaps also polar angle ($\delta\varphi$) as a means of relative position. A number of the previously mentioned works have in fact considered angular and lateral offsets of the sensors, be it implicitly via adaptive filters [158, 156] or explicitly via transformations between coordinate systems [142]. Arguably, specific transformations might be useful in cases where relative orientation and location of the device with respect to the body were known. Indeed, some devices are apt to be worn at specific locations, such as watches, headphones or glasses [178]. Without doubt, the situation is much more complex for handhelds and particularly so for mobile phones, for which, as opposed to wearables, the form-factor does not automatically imply where or how the device is worn or carried, especially when the device is not in use [150].

The precise location is furthermore influenced by clothing and what else is carried along, and varies according to the social and physical context of the user. Ichikawa et al. [150] conducted a thorough study by means of contextual interviews with a fixed set of questions. To eliminate subjective views as much as possible, the interviews were conducted in the public. People were interviewed in Helsinki, Milan and New York. Busy places were excluded from the studies as people's behaviour is supposedly biased under corresponding circumstances. A total of 225 males and 194 females were interviewed, of which 67 were less than twenty years old, 192 were of age between twenty and twenty-nine, 110 between thirty and forty-nine, and 48 were over fifty. In 34% of all cases, phones were carried in the trousers pockets, followed by shoulder bags with 33%. Interestingly, New York citizens differed significantly from others in that 67 out of 419 participants wore their phone in the trousers pockets, whereas only 36 respective 41 did so in Helsinki and Milan. Likewise, only 2 people from Helsinki carried their phones in belt enhancements as opposed to 18 and 15 people from Milan and New York. Whereas the study provides detailed statistics about whether the phone is worn in the back or front pocket, of which the front pocket is the clear "winner", it does not consider the phone's orientation within pockets or bags, i. e. whether the screen faces away from the body, or which way is up. 93% of the people however confirmed that the location where they carried the phone when they were interviewed was precisely the one where they would usually carry their phone [150]. For the remainder it was reported that people would wear their phones at different locations because they expected calls, or simply due to different clothing. Part of the interviewees also mentioned that they would sometimes store their phones elsewhere so as to avoid being disturbed by incoming calls. Lastly, the study also revealed huge differences in the wearing preferences among men and women. In 57% of the cases, men would prefer their trousers pockets in contrast to 66% of the women favouring their shoulder bags. The principal locations also turned out to vary with age. Below thirty, most people (40%) carry their phones in the trousers pockets, followed by shoulder bags (35%) and other locations. People over thirty supposedly prefer shoulder bags (28%) over trousers pockets (25%), as well as belt enhancements or upper body pockets over other bags.

A number of studies were eventually concerned with actually determining the mobile phone's "context" or relative location and orientation with respect to the body. In this regard, [285, 284, 109] describe a sensor board including photodiodes, accelerometers, barometer, thermometer, microphone and a few other entities, which was built into a mobile phone and then used to deduce the phone's context, more precisely whether the phone was being held in the user's hand, lay on a table, or was located in a suitcase. Kunze et al. infer relative device location from motion patterns [178]. They argue that motion patterns are rather specific to the location where a device is worn, be it for instance in a trousers pocket or in a hand, and that walking can be easily recognized regardless of device location and orientation. Also, potential movements might be constrained at certain locations, such as e. g. tilting the head about more than 90°. This view is largely theoretical and the authors name a few situations where such assumptions will not hold (when seen from a sensor's perspective), for example if a person were in the process of lying down. Therefore, and because they deem that walking is easily recognized, their method is restricted to phases where the user walks. In their experiments, sensors were placed at either the wirst, the head, the left trousers pocket, or the left breast pocket. Features were computed from the distributions of rate of turn, acceleration and magnetic field sensing, using a one-second sliding window with a half-second overlap. According to their results, they were able to predict on-body location with an accuracy of about 80% when algorithmically detecting walking patterns, or 90% when the frames had been previously labeled accordingly [178]. In a later work, Kunze et al. [179] then determine the relative orientation of the device with respect to the body, provided that the user is walking. For this, they first determine the gravity vector and thus the horizontal plane, and subsequently project all three-axis acceleration measurements onto that plane. The principal component of the latter then points into the direction in which the user is walking. On the downside, however, this does only relate the device to the body, but does not allow for further alignment within a global reference frame between multiple devices. Shi et al. [296] used low-cost gyroscopes and accelerometers to infer radius and angular velocity, based on the notion of specific motion of a limb around a joint in a rigid body model. They demonstrate that their algorithm is invariant to orientation as it only uses the magnitude of the vectorial measurements. In comparison to Kunze et al. they use a significantly larger window of ten seconds for which they compute the sample distribution's mean, variance, kurtosis, skewness and characteristic quartiles. Their results show that they could predict device location with about 91% accuracy from different recordings, each of ten minutes length and with no instructions regarding device orientation given to the experiment's participants. It should be noted, however, that only four individuals participated in their experiments. In a slightly larger study [331], Vahdatpour et al. asked 25 participants to attach sensors to their bodies inside predetermined regions, but without further instructions on the precise locations or modes, e. g. whether the sensors ought to be attached to the skin or clothing (figure 38). Similar to [178] and [296], they conclude that walking were the predominant activity during the day and therefore their method would be mainly based on walking patterns. Aside from phases where the users walk, they also consider "general activity". In addition to features in the time domain, they compute features in the frequency domain, for instance with respect to energy because they assume that e. g. sensors closer to the foot experience stronger impulses than those farther away. For phases of "general activity", however, they are primarily interested in changes of orientation over time. Like [296] they use accumulated or maximum values from the features which were computed for each

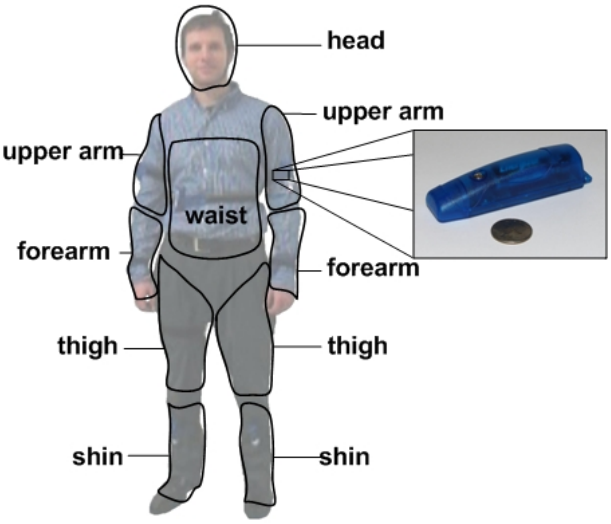


Figure 38.: Predetermined regions for sensor placement. Figure taken from [331].

axis of the sensors to ensure that their method is invariant under rotation. In a first step, their method performed better than [178] in regard of detecting walking patterns, and in a second step, prediction of location was reported to be near perfect when classification was done on the basis of per-user training data. Interestingly enough, they are the first to state that this would likely not apply to real life scenarios where personalized training data are not guaranteed to be available. Nonetheless, a final evaluation of the classifier based on training data randomly drawn from the whole set of participants exhibits about 89% accuracy and very satisfying precision and recall for each location.

3.1.4 *Discussion*

The survey of related work has shown that a variety of methods allow for more or less accurate estimation of device orientation and location. Some studies have also considered angular and lateral offsets of the sensors from the body. The fusion of three-axis accelerometers, magnetometers and gyroscopes allows for highly accurate attitude estimates, and the remaining uncertainty can be reduced effectively through well-studied mechanisms like Kalman or particle filters. Such filters not only help to reduce sensor noise or the effects caused by drift, but also smooth sudden acceleration or rotation, which is likely experienced in a scenario where a mobile agent is carried in a hand, a bag, temporarily stuffed away, or perhaps even during sports. Whereas filter design can be arbitrarily complex, the resulting models are often linear or, for instance in case of the EKF, linear approximates to non-linear transformations in terms of Taylor series expansion, and can therefore easily be employed in realtime applications. Also, nearly all modern smart phones feature a minimum set of the aforementioned types of sensors, are capable of sampling at reasonably high frequencies, and provide sufficiently high resolutions for quantization of the raw signals. However, using numerous sensors simultaneously, possibly at high sampling rates and for longer periods of time, is certainly an issue (e. g. in terms of trading power for for battery life) that will have to be further explored. Constandache et al., for instance, have investigated the trade-off between a given energy budget and localization accuracy [60], whereas Priyantha et al. propose that sensing and processing should be offloaded to additional dedicated processors [252]. In yet another work on orientation estimation [38], Blanke and Schiele completely avoided the use of gyroscopes in favor of reduced energy consumption. Instead, they sampled accelerometers and magnetometers at merely 50 Hz. The samples were then smoothed with a KF in conjunction with an adaptive noise model for magnetic disturbances and motion. Depending on the on-body locations of the sensors, their system revealed maximum errors between 7° and 27° in comparison to a commercially available high performance system.

One major issue of orientation estimates in the context of social signal processing, or more precisely social interaction geometry, is the problem of relating sensor attitude (or location) to the user's body or any global reference frame. Some of the related works have done so through adaptive filtering, while others have proposed static transformations which may be selected depending on the presumed on-body location of the device. Several of the previously stated studies have shown that a device's on-body location can in fact be predicted with high accuracy. In addition to that, others have shown that the most popular locations where people wear their mobile phones are trouser pockets and shoulder bags, followed by only a few other locations, albeit much less frequent. It follows that if one were to learn specific models for each reasonable location, the necessary number of distinct models would be rather low. What still has to be accounted for, though, is the variety of orientations and, if things were taken to the extreme, also whether devices were carried inside protecting cases or sleeves. Also, sensor performance might vary depending on the specific type of surrounding clothing or textiles. At the bottom line, computing specific models for each of the most popular locations, in conjunction with varying models depending on possible and location-dependent orientations, is without doubt realistic and practical, whereas incorporating additional means of dealing with protective cases or textiles is most certainly not due to the multitude of options and the required vast number of potential models. The insignificance of the latter is corroborated by Maurer et al. [207] who evaluated various sensor locations during the recognition of basic activities such as walking, sitting, running, or standing, and found that sensors worn in pockets perform only slightly less than those placed in bags.

Next to orientation and on-body location, some of the related work was concerned with (mainly indoor) localization of devices with respect to some global reference frame. In regard of social interaction geometry, it does not matter whether position in a global reference frames is expressed in latitude/longitude or arbitrary units, and also not whether the frame is related to earth, as long as metrics exist that allow for accurate relation of multiple devices to each other. Pedestrian DR and, in its more general form, IN are well-studied techniques which are already employed in present-day smartphone scenarios. Using sensor fusion and sophisticated filter designs, these techniques allow for accuracies of about 1.5 meters. Nevertheless, since both DR and IN are prone to quick accumulation of errors, countermeasures have to be taken. The NavShoe [98], for instance, feeds pseudo-measurements of zero velocity into the corresponding filter during stance phases, i. e. during those periods when the foot is firmly placed on the ground as the user is walking, therefore effectively keeping prediction errors within reasonable limits. On the other hand, further data sources can be incorporated for localization, even if that would mean using external infrastructure such as GPS or WiFi beacons. The latter are likely not available in many outdoor and particularly indoor scenarios. Other than that, there is no good cause against using *supplemental* infrastructure-dependent mechanisms to improve localization quality.

Most time-of-flight-, fingerprinting-, pedestrian DR- or IN-based techniques do nonetheless provide only coarse location estimates at a level that is insufficient for applications of social interaction geometry models. The distribution of δd in the data from both the first and second experiments (figures 8e, 11, 28c, and 29c), as well as the corresponding shapes of the Gaussians in the final models, are relatively “broad”, so to speak, therefore implying that, in particular, centimeter-level accuracy is not mandatory. Still, average measurement errors of one meter or more are effectively too much.
From the evaluation of the algorithmic model for detection of social interaction geometry (sections 2.3.5 and 2.4.3), it is evident that the most important variables in dyadic interaction are mutual distance (δd) and the relation of the shoulder lines ($\delta\theta$). The relative position in terms of the polar angle ($\delta\varphi$) naturally refines the model and conveys more information to the classifier, but it correlates with both shoulder orientation and distance to a large degree, and is also much harder to determine, since either precise knowledge of the absolute position or general means of trilaterion are required. If $\delta\varphi$ were to be left out, that would leave the model with shoulder orientation and distance only. The latter is especially interesting as, despite of the aforementioned, too inaccurate localization techniques, ultrasound-based methods like the *BeepBeep* framework proposed by [229] are in fact capable of highly accurate estimates of distance between mobile devices, independent of any external infrastructure or explicit time synchronization. BeepBeep and related methods were further discussed in [95] to the point of application among more than just two devices, although that particular discussion seemed to be largely theoretical.

The remainder of this chapter will therefore present alternatives for estimations of orientation and location of the user based on their mobile phone. Section 3.2 shows a system for the estimation of the user’s body attitude from device attitude. The proposed system is based on a linear regression model from various sensor signals to the orientation of the shoulders about the yaw-axis, as well as the angle between the torso and the leg at which the mobile device is worn, the latter of which comes as a by-product and might e. g. be useful in scenarios that benefit from information such as whether a person is standing, sitting or walking. Nevertheless, relative orientation $\delta\theta$ about the yaw-axis is the only relevant attitude measure for application in social interaction geometry. Section 3.3 subsequently presents an ultrasound-based system for measuring distance among multiple users. The proposed system makes use of a mobile array of conventional ultrasound sensors in order to estimate interpersonal distances δd at reasonable accuracy for the proposed interaction geometry model. In addition to the mandatory distance measure, the system can furthermore aid in the estimation of relative position in terms of the polar angle $\delta\varphi$. Readers should note that a similar system was proposed in [205], based on a preliminary version of the model presented in chapter 3, previously published by Groh et al. in [123]. In [205], Matic et al. estimate relative distance and orientation through WiFi- and a logical orientation sensor provided by the Android platform. In addition to interaction geometry, their model includes a binary variable indicating speech activity as further evidence for social interaction. For this, accelerometers were strapped around the subjects' chests. For their WiFi-based distance estimates, they report a 50% percentile of 0.5m measurement error when using the same phone model, and up to 1.8m when using different models. Both on-body location and orientation of the devices were controlled and constant parameters throughout their experiments, avoiding the necessity of dynamically relating device location and orientation to the user's body. Interestingly, the authors suggest to measure the *stability* of orientational arrangements in terms of the standard deviation of relative orientation. It is briefly mentioned that using this feature instead of absolute measures could furthermore render the transformation between device and body orientation obsolete. Without doubt, their proposed feature contributes to models of interaction geometry as additional means of robustness. Solely relying on that feature, however, would lead to a loss of important information. For example, pairwise orientation is of course also stable if one person were to face the back of another. In such and similar cases, using only the proposed feature together with interpersonal distance is clearly not sufficient for the distinction of interaction from non-interaction. Second, akin to the fact that measured distance correlates with social distance [67], similar holds for orientation and the affective meaning of a social situation, subsequently shown by the same first and second authors in [206].

3.2 A SYSTEM FOR MEASURING PERSONAL HEADING

Previous studies have explicitly defined static transformations from device to body orientation in terms of rotations dependent on the specific location where the device is worn [142], or implicitly incorporated such transformations by means of adaptive filtering [158, 156]. Others have related device attitude to the direction into which the user was heading via PCA of the projection of three-axis acceleration measurements onto the horizontal plane while the user was walking [179]. In the context of social interaction geometry, the most relevant information about a user's orientation is without doubt given in terms of his or her heading, as the difference between the two distinct headings in a dyad yield people's relative orientation towards each other. Instead of manually defining one or more such transformations, a time-invariant *general linear model* [201, 34] is proposed which relates a number of sensor measurements and derived features to the user's heading and the angle between the torso and the leg at which the device is worn. Note that the latter is not used in the context of social interaction geometry, but comes easily as a by-product in the overall process. The system fusions mobile phone accelerometers, magnetometers and gyroscopes, and is therefore able to estimate the user's heading with respect to a global ENU reference frame, as opposed to the local reference frame commonly found in activity recognition tasks. Similar to the works of Schwarz [289] or Roetenberg [272], who combined inertial sensors with methods from Computer Vision (CV) for estimating different postures of the human body, the model is trained from a dataset which provides the sensor measurements along with the corresponding ground truth. This dataset was acquired by combining measurements from mobile phone sensors with body posture estimates from a Microsoft Kinect system. The final model, in either personalized or general form, works independent of CV systems such as the Kinect and will be shown to have sufficient accuracy for algorithmic models of social interaction geometry, such as the one presented in section 2.3. Note that the model and the dataset were created in the proceedings of [69].

3.2.1 *How the Kinect works*

According to Microsoft, the Kinect was built to "revolutionize the way people play games and how they experience entertainment", enabling "people to interact with the games through their body in a natural way" [357]. Aside from gaming, though, the system has been adopted by computer science, robotics, and various other fields, for example as a means of altitude control for helicopters [314], for three-dimensional object manipulation on a desktop display [259], or for human pose estimation as will be discussed below.

Before the advent of the Kinect and related techniques, body-pose was for instance estimated in several phases [108]. If possible, the current state was first extrapolated from previous state(s), as it was "deemed more stable to do the prediction at a high level (state-space) than at a low level (image-space)" [108]. Backtransformation from state-space to image-space would identify the relevant parts of the image, from which features would then be extracted, and finally the new state would be estimated according to the segmented image [108]. Setups of multiple cameras, or alternatively monocular image sequences, were common for three-dimensional pose estimation, and skeletal models of the body were used to incorporate domain knowledge such as the length of the limbs or the degrees of freedom of the joints [214]. On top of these skeletal models, volumetric and non-volumetric flesh models helped in relating state-space and image-space. Many approaches required the full visibility of at least the face and upper body, and were apt to experience problems as soon as parts of the body were occluded or cut off from the available image region [214, 164]. The "loss of depth and limb labeling information would furthermore make the "recovery of 3D pose [...] ambigious" [9]. Later works have combined computer vision and inertial sensing for further improvement of estimating limb position and orientation [245, 246].

The Kinect platform [1] consists of an infrared laser, infrared sensor, an RBG camera, a single tilt motor and four microphones. Through these it is capable of full-body three-dimensional motion capture, facial recognition, and voice recognition [357]. As a matter of fact, technical details have never been made available to the public by Microsoft, but were reverse engineered and correlate with a number of patents of PrimeSense, the company behind the system design [112]. In addition to the RBG image from the regular camera,

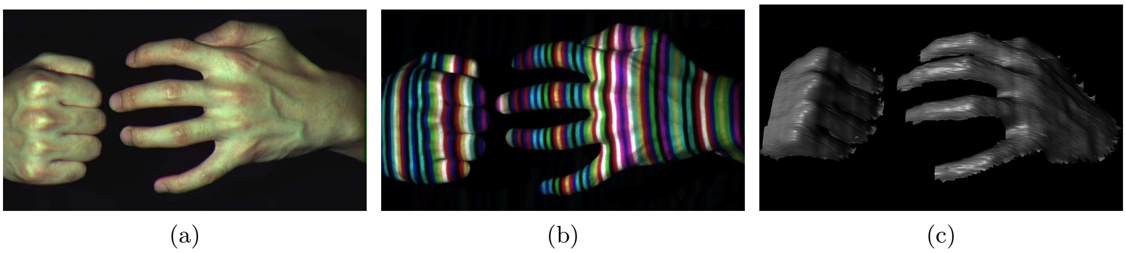


Figure 39.: Structured light principle. Figure taken from [356].

the system provides a depth map. Only the infrared laser and sensor are used for three-dimensional image reconstruction, based on the principle of *structured light* [356]. The structured light principle works by projecting a coloured or otherwise encoded pattern onto the scene from one point of view and capturing it from another. The relative distance between a particular point in the image plane as seen from the projector and the same point in the image plane as seen from the camera is inversely related to its depth. For each pixel its depth can hence be reconstructed, provided that calibration data for the projector and the camera are available, along with the details of the used pattern. The principle is illustrated in figure 39. Here the Kinect uses its infrared laser to project a pattern consisting of dots of varying size and spacing which is then captured by the infrared sensors. Both sensors are laterally displaced from each other, hence accounting for distinct points of view. The captured pattern is compared against a reference pattern, for which the system was calibrated at a plane at precisely known distance. All in all the system is accurate within one or two centimeters [168, 169], for which the "depth from stereo", i. e. the structured light principle, is further augmented by "depth from focus". The latter stems from the fact that objects at greater distances are perceived as more blurry. For this the system further incorporates an astigmatic lense with different focal lengths along the horizontal and vertical axes, so that a (theoretically) projected circle would appear as an ellipse "whose orientation depends on depth" [196].

The pose of the whole body is estimated from the computed depth map. After foreground segmentation, a *randomized decision forest* is used to predict which pixel of the depth map belongs to which part of the body, for which a total of 31 body parts are considered [298]. Each decision tree yields a probability distribution over the pixels for a specific body part. The modes of these distributions are subsequently computed via mean shift estimation [101]. Once the positions of the limbs, or more generally the body parts, have thus been determined, a set of candidate joints is then predicted [113, 298]. This process is illustrated in figure 40. The very high accuracy of the model is due to the fact that it has been computed on a huge dataset. This dataset is comprised of about 100,000 segmented and annotated images which, in addition, have been synthetically altered according to fifteen base characters, considering "both male and female, from child to adult, short to tall, and thin to fat" [196, 298]. Furthermore, height and weight are varied at random by ±10%. Each distinct pose is mirrored to prevent one-sided bias to the left or right. Consequently,

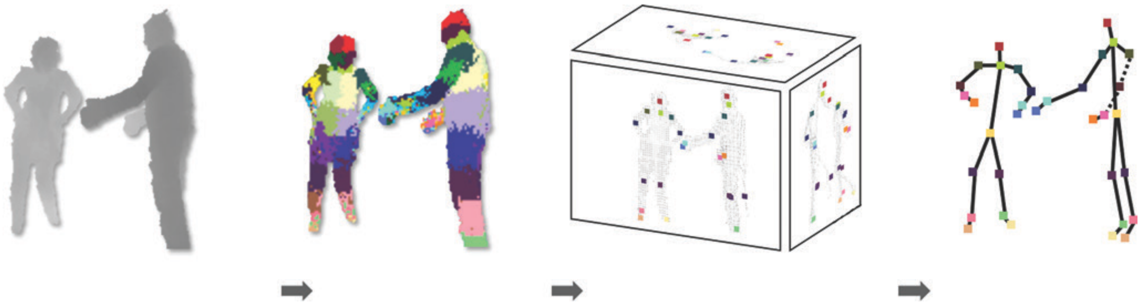


Figure 40.: Candidate joints prediction. Figure taken from [298].

the final dataset consists of millions of samples.
As the Kinect works on a frame-by-frame basis, no spatio-temporal tracking is necessary, although it could improve the predictions [196]. The system proposed by [298] is capable of processing at 5 Hz. However, the Kinect is subject to hardware and software delay, still resulting in about 30 Hz, which is relevant with respect to the further proceedings of this section. In regard of general-activity poses, i. e. those not subject to prior constraints regarding the range of motions, improvements have been reported concerning e. g. occlusions of body parts, acknowledging the fact that joints are inside the body whereas segmentation is done on the surface [113], or introducing so-called metric space information gain in order to optimize entropy of the probability distributions in metric space [247].

3.2.2 *A model for linear regression*

Linear regression models predict the values of one or more response variables t, given the values of one or more regressor variables x, for which the models need only be linear in their parameters but not necessarily the input [34, 218]. For a single scalar response variable, the model is generally defined as

$$y(\mathbf{x}, \boldsymbol{\beta}) = \beta_0 + \sum_{i=1}^{M-1} \beta_i \sigma_i(\mathbf{x}) \ , \tag{110}$$

where $\mathbf{x} = (x_1, \ldots, x_D)^T$ is a vector of the input variables, and β and σ denote the set of M model parameters and $M-1$ basis functions. Letting $\sigma_0(\mathbf{x}) = 1$, the model can be conveniently rewritten as

$$y(\mathbf{x}, \boldsymbol{\beta}) = \sum_{i=0}^{M-1} \beta_i \sigma_i(\mathbf{x}) = \boldsymbol{\beta}^T \boldsymbol{\sigma}(\mathbf{x}) \tag{111}$$

with $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{M-1})$ and $\boldsymbol{\sigma}(\mathbf{x}) = (\sigma_0(\mathbf{x}), \ldots, \sigma_{M-1}(\mathbf{x}))$.

The proposed model considers two target variables, namely the user's heading and angle between torso and leg. The basis functions simply correspond to identity, so that $\sigma(x) = [1, x]^T$. The model is therefore a *multiple linear regression model* of the form

$$\mathbf{t} = y(x, \mathbf{B}) + \boldsymbol{\epsilon} = \mathbf{B}^T x + \boldsymbol{\epsilon} \, , \tag{112}$$

for which the parameters have been arranged in the matrix $\mathbf{B}$ and $\boldsymbol{\epsilon}$ models the statistical error. Note that $\boldsymbol{\epsilon}$ is supposed to follow a normal distribution. This is particularly justifiable in the context of sensor measurements because they can be regarded as the sum of multiple random variables, i. e. the measured entities themselves plus systematic and random errors from numerous hardware and software sources. The normal distribution then follows from the central limit theorem [188].
Estimation of the model parameters is straightforward, e. g. via gradient descent. Due to the linearity of both the parameters and the input, the values of each of the response variables in equation 112 correspond to points on a hyperplane. Therefore, a common choice of loss function is *squared loss*. Given a matrix $\mathbf{X} \in \mathbb{R}^{N \times D}$, whose rows correspond to a set of N vectors of D independent regressor variables, along with a matrix $\mathbf{Y} \in \mathbb{R}^{N \times K}$, whose rows determine the corresponding K-dimensional responses, the loss function f is

$$f = \frac{1}{2} (\mathbf{XB} - \mathbf{Y})^2 \; . \tag{113}$$

Differentation with respect to $\mathbf{B}$ leads to the closed-form solution

$$\frac{df}{d\mathbf{B}} = \mathbf{X}^T(\mathbf{XB} - \mathbf{Y}) \overset{!}{=} 0 \;\Leftrightarrow\; \mathbf{X}^T\mathbf{XB} = \mathbf{X}^T\mathbf{Y} \;\Leftrightarrow\; \mathbf{B} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \, , \tag{114}$$

given by the Moore-Penrose pseudo-inverse of $\mathbf{X}$. This form allows for the estimation of the model parameters also for underdetermined systems, i. e. when $\mathbf{X}$ is not of full rank. Note that, from a probabilistical perspective, linear regression is equivalent to modeling the predictive distribution $p(\mathbf{t}|\mathbf{x})$. It can be shown that the squared loss provides an optimal solution through the conditional expectation of $\mathbf{t}$ [34], or equivalently that it is the best linear unbiased estimator of the model parameters [242].

As already indicated at the beginning of this section, the proposed system should predict the user's heading with respect to the world ENU reference frame, as well as the angle between torso and leg. It is important, though, that the underlying model itself is decoupled from absolute values such as the heading. For this, the model responds with the difference Δh_{bd} between the device heading h_d and the body heading h_b, the latter of which is defined orthogonal to the user's shoulder line. Let $x = (h_d, x_2, \ldots, x_D)$ a vector of values of the regressors, i. e. the device heading and additional features $x_2 \ldots x_D$. Then the two variables, under a linear regression model, are related as

$$h_b \sim \beta_0 + \beta_1 h_d + \beta_2 x_2 + \ldots + \beta_D x_D \, , \tag{115}$$

from which it follows that

$$h_b - \beta_1 h_d \sim \beta_0 + \beta_2 x_2 + \ldots + \beta_D x_D \; . \tag{116}$$

The term on the left-hand side thus describes the difference between the two headings if, and only if, $\beta_1 = 1$. This is however an elementary assumption of the proposed model, meaning that the location and orientation of the mobile phone are supposed to be *invariant* with respect to the body. As a consequence, the differential heading Δh_{bd} is henceforth defined as:

$$\Delta h_{bd} = h_b - h_d \tag{117}$$

It follows that one can easily add h_d to Δh_{bd} so as to predict the absolute heading. Note that using the differential instead of the absolute heading not only makes the model invariant to absolute orientation. The corresponding removal of h_d from the right-hand side of the equation also yields a major simplification of the model, as it helps to avoid the mandatory specific treatment of h_b and h_d due to their periodicity.

3.2.3 *The dataset*

A new dataset was compiled from measurements of the stationary Kinect and a mobile phone in a series of six experiments, each of which was conducted with another subject. The subjects carried the phone in their left or right trousers pocket in various orientations. Great care was taken so as to avoid occlusions of the body parts as well as possible interferences. A custom software system was used to record and process the datastreams from both devices [69]. The system is comprised of two subsystems, one on a personal computer with a cable-connection to the Kinect, and one a mobile phone. The measurements from the Kinect as well as those from the mobile phone sensors were acquired using Microsoft's Kinect and Windows Phone 7 software development kits [63, 64]. For this, the mobile phone application sent a continuous stream of data to the personal computer via TCP/IP networking. Both datastreams were then aligned and multiplexed in order to correct for the corresponding delays, the latter of which had been thoroughly determined in a series of recordings prior to the actual experimental sessions. These more or less systematic delays are caused by the long chain of hardware and software components. For the Kinect, the average delay was determined to be 70 ms as opposed to 150 ms for the mobile phone. Consequently, the Kinect measurements were buffered and multiplexed with those data arriving from the phone after an explicit delay of 80 ms. In accordance with section 3.2.1 it was furthermore determined that the Kinect provides new data about every 32 ms with a standard deviation of 1.2 ms, which is why a sampling rate of 31 Hz was chosen for sampling from the Kinect. The mobile phone, on the other hand, is capable of providing a constant stream of new measurements at 20 Hz. In order to avoid sophisticated frequency harmonisation, e. g. in terms of up- and subsequent downsampling [359], or spherical interpolation for quaternions [297] (refer to section 2.2.2.2), the respective least recently received data from the devices were processed.

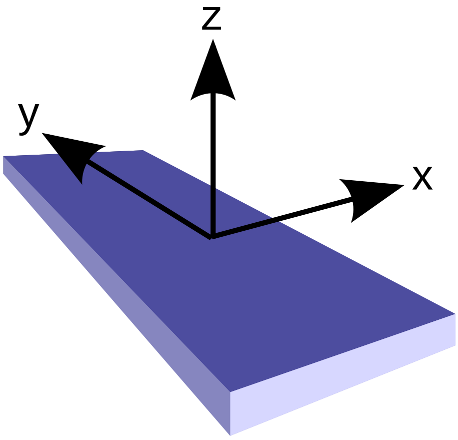


Figure 41.: Windows Phone™ coordinate system.

3.2.3.1 *Postprocessing*

Various features were computed during postprocessing of the newly acquired dataset. The basic features are given by the raw values of the three-axes sensor measurements of the gyroscope, accelerometer, and magnetometer. As for the heading, recall that although the underlying regression model will respond with the differential heading Δh_{bd}, the absolute headings h_d of the device and h_b of the body have to be known for training. The orientation of the device is computed and provided by the mobile phone's firmware, for which it integrates and filters accelerometers, magnetometers and gyroscopes, expressed as a rotation quaternion $\mathbf{q}$. This rotation quaternion describes the orientation of the axes of the device's local coordinate system in the world ENU reference frame. According to the SDK [64], the axes of the phone are laid out as depicted in figure 41. For the set of features, h_d is computed from this rotation quaternion. Generally speaking, the device heading has to be defined in relation to a particular entity such as e. g. the phone's y-axis, z-axis, or the intersection of its y/z-plane with the global x/y-plane of the world's ENU reference frame. In the context of the model, this choice is arbitrary provided that the reference is the same throughout the whole process. However, if only one of the phone's x-, y- or z-axes were chosen specifically, that would lead to singularities whenever the phone were oriented such that this axis were parallel or even very close to the global z-axis, in other words the gravity vector. In order to avoid these singularities, h_d is determined as a weighted sum of the angles between the global ENU y-axis and the phone's y- and z-axes, respectively. As is known from section 2.3.2.1, in general the circular mean and arithmetic mean tend to differ. Likewise, the weighted sum has to take into account the circular properties of the aforementioned angles. Therefore, let

$$\mathbf{y} = \mathbf{q}(0 + 0i + 1j + 0k)\mathbf{q}^* = 0 + y_x i + y_y j + y_z k \tag{118}$$

$$\mathbf{z} = \mathbf{q}(0 + 0i + 0j - 1k)\mathbf{q}^* = 0 + z_x i + z_y j + z_z k \tag{119}$$

for the rotation quaternion q and the quaternion product as discussed in section 2.2.2.2. It follows that $||\mathbf{y}|| + ||\mathbf{z}|| = 1$ which is then also true for the projections $\mathbf{y'} = (y_x, y_y)^T$ and $\mathbf{z'} = (z_x, z_y)^T$ onto the x/y-plane. Now let

$$\mathbf{r} = (r_x, r_y)^T = (\alpha y_x + (1-\alpha)z_x, \alpha y_y + (1-\alpha)z_y)^T \,, \tag{120}$$

where $\alpha = ||\mathbf{y}|| = 1 - ||\mathbf{z}||$. This would allow for a preliminary definition of h_d:

$$h_d = \operatorname{arctan2}(r_y, r_x) = \operatorname{arctan2}\left(\alpha y_y + (1-\alpha)z_y, \alpha y_x + (1-\alpha)z_x\right) \tag{121}$$

The quality of the result is further increased by applying a logistic function of the form

$$f(\alpha) = \frac{1}{1 + e^{-\lambda(\alpha - \frac{1}{2})}} \,. \tag{122}$$

Depending on the choice of λ this function will help to attenuate either one of $\mathbf{y}$ or $\mathbf{z}$, considering their respective length. Thus h_d is finally defined as

$$h_d = \operatorname{arctan2}\left(f(\alpha) \cdot y_y + (1 - f(\alpha)) \cdot z_y, f(\alpha) \cdot y_x + (1 - f(\alpha)) \cdot z_x\right) \,, \tag{123}$$

for which, in this particular context, $\lambda = 16$ has empirically proven to yield the best results.

Now that the device heading has been defined in relation to both the phone's y- and z-axes, the device attitude needs to be adapted accordingly to yield heading-invariant attitude information. This adaption is done by transforming the device's rotation quaternion $\mathbf{q}$ by an inverse rotation of h_d about the z-axis, yielding the updated quaternion $\mathbf{q'}$. Rewriting the quaternion rotation operator in matrix notation (as in equation 5) yields the DCM

$$2 \cdot \begin{pmatrix} q'^2_0 + q'^2_1 - \frac{1}{2} & q'_1 q'_2 - q'_0 q'_3 & q'_1 q'_3 + q'_0 q'_2 \\ q'_1 q'_2 + q'_0 q'_3 & q'^2_0 + q'^2_2 - \frac{1}{2} & q'_2 q'_3 - q'_0 q'_1 \\ q'_1 q'_3 - q'_0 q'_2 & q'_2 q'_3 + q'_0 q'_1 & q'^2_0 + q'^2_3 - \frac{1}{2} \end{pmatrix} \tag{124}$$

which is equivalent to the following DCM based on a yaw/pitch/roll rotation sequence in terms of Euler angles ϕ, θ, ψ:

$$\begin{pmatrix} \cos\psi\cos\phi + \sin\psi\sin\theta\sin\phi & -\cos\psi\sin\phi + \sin\psi\sin\theta\cos\phi & \sin\psi\cos\theta \\ \cos\theta\sin\phi & \cos\theta\cos\phi & -\sin\theta \\ -\sin\psi\cos\phi + \cos\psi\sin\theta\sin\phi & \sin\psi\sin\phi + \cos\psi\sin\theta\cos\phi & \cos\psi\cos\theta \end{pmatrix} \tag{125}$$

Let R_{ij} reference the element at row i and column j of the above matrix. Then yaw (ϕ), pitch (θ) and roll (ψ) of the heading-invariant device attitude adhere to

$$\phi = \operatorname{sgn}\left(\sin^{-1}\frac{R_{21}}{\cos\theta}\right)\cos^{-1}\frac{R_{22}}{\cos\theta} \,, \tag{126}$$

$$\theta = \sin^{-1}(-R_{23}) \quad \text{and} \tag{127}$$

$$\psi = \operatorname{sgn}\left(\sin^{-1}\frac{R_{13}}{\cos\theta}\right)\cos^{-1}\frac{R_{33}}{\cos\theta} \,. \tag{128}$$

Together, h_d and h_b serve to determine the differential heading and hence the corresponding response variable Δh_{bd}, whereas the three Euler angles ϕ, θ and ψ serve as regressor variables. A number of additional features moreover yield temporal information like the mean, the variance (as in energy), and the Pearson correlation coefficients for the angles respective pairwise angles over the past second. These temporal features might lead to the question as to what extent the irrefutable correlation between temporal adjacent samples constitutes a problem since typical machine learning models assume i.i.d. samples [76]. Many models however perform quite well in spite of erroneously (and knowingly) assumed independence, for instance Naïve Bayes, instead of exploiting e. g. sequential data. Also, even if there were no such features like these that correspond to shifting windows, there is likely always an underlying physical dependency between subsequent samples. For this dataset, the short (one second) temporal correlation of the samples is considered insignificant in relation to the size of the dataset. Furthermore, the order of the samples is randomized and only a subset of the data will be used during training.

Three more features were added for a rough assessment of periodicity in the movements as reflected in the Euler angles. Walking and running, for example, are expected to show up with different base frequencies in one or more of the signals ϕ, θ and ψ. The step frequency of humans usually lies well down below 200 steps per minute [48], which is equivalent to a maximum frequency of about 5 Hz. The sampling rate of the sensors lies well above the Nyquist frequency of 10 Hz. With the chosen sampling rate of 20 Hz, a corresponding Short-Term Fourier Transform (STFT) yields the amplitudes (and phases) of $\frac{N}{2}$ discrete frequencies of N bins of the windowed input signals, ranging from 0 Hz to $\lfloor\frac{N}{2}\rfloor \cdot \frac{20\ \mathrm{Hz}}{N}$ in equidistant intervals [359]. That frequency which corresponds to the largest amplitude is then selected as the feature value for ϕ, θ and ψ, respectively.

Finally, α_{lt} is determined as the angle between the torso and the leg corresponding to the trousers pocket in which the device is worn. For this, the axis of rotation is defined as a line through both hip joints. This line corresponds to the intersection of a plane through the hip joints and the center of the shoulder joints with another plane through the hip joints and the knee of the respective leg. α_{lt} therefore corresponds to the angle between the front-facing normals of these planes. Despite the arguable limits of human motion, the angle is defined on the whole interval $[0, 2\pi]$, for which e. g. π corresponds to a setting where a person were lying flat on the back. The angle is hence defined as

$$\alpha_{lt} = \pi + \alpha_{sgn} \cos^{-1}(n_{hs} \cdot n_{hk}) \;, \tag{129}$$

where n_{hs} and n_{hk} denote to the normals of the planes between hip and shoulder or knee, respectively, and α_{sgn} is either $+1$ or -1, depending on the direction from which an observer looks at the body. As a reference, the observer is assumed to be located to the left of the body, so that $\alpha_{sgn} = +1$ if, and only if, $n_{hs} \times n_{hk}$ points away from the observer.

3.2.4 *Evaluation*

Due to the fact that both response variables Δh_{bd} and α_{lt} are independent of each other, they were each modeled and evaluated individually. For those features related to STFT, Pearson correlation, mean or variance, appropriate window sizes were chosen, and for each model the best subset of regressor variables was determined. The results were then verified via 10-fold cross-validation. The final set of regressor variables for predicting Δh_{bd} is given by the yaw (ϕ), pitch (θ) and roll (ψ) angles, along with the corresponding temporal features, namely the standard deviations and the Pearson correlation coefficients:

$$\Delta h_{bd} \sim \left(\phi, \theta, \psi, \sigma_\phi, \sigma_\theta, \sigma_\gamma, \rho_{\phi\theta}, \rho_{\phi\psi}, \rho_{\theta\psi}\right)^T \theta + \epsilon \tag{130}$$

Somewhat unexpected, α_{lt} is also best modeled by the same set of regressors:

$$\alpha_{lt} \sim \left(\phi, \theta, \psi, \sigma_\phi, \sigma_\theta, \sigma_\gamma, \rho_{\phi\theta}, \rho_{\phi\psi}, \rho_{\theta\psi}\right)^T \theta + \epsilon \tag{131}$$

The final feature sets were determined as follows: The original set of features was first partitioned into thirteen equivalence classes such as angles, raw measurements from the accelerometers or magnetometers, related means, standard deviations, correlations, etc. Denote this set of feature groups as $\mathcal{F}$. The final set of features was then selected by cross-validating all models arising from the elements of the powerset $2^{\mathcal{F}} \setminus \emptyset$. All the same, the window sizes for the temporal features were varied between a half and three seconds in intervals of a half second. The regressors were selected based on the comparison of the R^2 respective *adjusted* R^2_{adj} scores. The R^2 score is defined as

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y_i})^2}{\sum_i (y_i - \bar{y_i})^2} \tag{132}$$

and quantifies which fraction of the variance of the data is explained by the variables as opposed to a model with constant mean [171]. As R^2 is monotonically increasing when new parameters are added to the model, R^2_{adj} is defined so as to compensate for an increase that might have been caused by chance:

$$R^2_{adj} = 1 - \frac{\sum_i (y_i - \hat{y_i})^2 / \mathrm{dof}_r}{\sum_i (y_i - \bar{y_i})^2 / \mathrm{dof}_t} , \tag{133}$$

The degrees of freedom $\mathrm{dof}_r = N - M - 1$ and $\mathrm{dof}_t = N - 1$ for N observations and M parameters account for the fact that both sums of squares (divided by N) are biased estimators of variance. Table 23 shows the values of these measures for the final sets of regressor variables. According to these results, the models explain about 65% and 85% of the observations' variance. The values of the adjusted measures are in fact very close to the normal scores, thereby indicating that all of the selected variables contribute to the model. This is further corroborated by the p-values under the null hypothesis that the regressor's coefficients were zero. Interestingly, the scores get much better (~ 92%) when the intercept term is removed from the model. On the other hand, this leads to a

	Δ_{hd}	α_{lt}
R^2	0.656	0.857
R^2_{adj}	0.645	0.852

Table 23.: Goodness of fit for the response variables Δ_{hd} and α_{lt} after 10-fold cross-validation.

	ϕ	θ	ψ	σ_ϕ	σ_θ	σ_ψ	$\rho_{\phi\theta}$	$\rho_{\phi\psi}$	$\rho_{\theta\psi}$
ϕ	1.00	-0.85	0.04	0.80	-0.78	0.17	0.80	0.50	-0.12
θ	-0.85	1.00	0.05	-0.70	0.84	-0.13	-0.63	-0.33	0.16
ψ	0.04	0.05	1.00	0.19	0.00	0.78	-0.05	0.13	0.87
σ_ϕ	0.80	-0.70	0.19	1.00	-0.76	0.21	0.61	0.47	0.09
σ_θ	-0.78	0.84	0.00	-0.76	1.00	-0.15	-0.76	-0.31	0.08
σ_ψ	0.17	-0.13	0.78	0.21	-0.15	1.00	0.15	0.33	0.41
$\rho_{\theta\phi}$	0.80	-0.63	-0.05	0.61	-0.76	0.15	1.00	0.54	-0.24
$\rho_{\psi\phi}$	0.50	-0.33	0.13	0.47	-0.31	0.33	0.54	1.00	-0.06
$\rho_{\theta\psi}$	-0.12	0.16	0.87	0.09	0.08	0.41	-0.24	-0.06	1.00

Table 24.: Pairwise correlation of regressor variables.

non-normal distribution of the residuals. Whether to remove the intercept term is thus a question of "usefulness" versus "correctness" of the model. It is worth mentioning that some of the regressors exhibit linear correlations (table 24). This is not surprising from a physical point of view, and also statistically speaking for the temporal features, for instance in cases where signals and their standard deviations are constant for some time. On a final note in regard of window size for the temporal features, the best results have been found for windows of one second, or about one and a half seconds (32 frames at 20 Hz) for the STFT-based features. The latter were however not included in the final models.

Next, analysis of the residuals should attribute to the "correctness" of the models. According to section 3.2.2, linear regression models assume that the values of the response variables correspond to points on a higher-dimensional manifold, in this particular case a hyperplane. The remaining statistical error ϵ is explained by Gaussian noise. In other words, $\epsilon \sim \mathcal{N}(0, \sigma^2)$ follows a normal distribution with zero mean and constant variance σ^2, or equivalently

$$y|x \sim \mathcal{N}(\beta^T x, \sigma^2)\,, \tag{134}$$

for which the mean of the *true* distribution of y, given x, is linearly increasing in x [34, 218, 184]. Recall that the residuals only serve as estimates, whereas mean and variance of the true distribution of the statistical error are generally unknown. It follows that the residuals have to be independent and adhere to a normal distribution with zero mean and constant variance [171]. As opposed to errors, residuals do not have constant variance, though. This is a consequence of the fact that observations gain more influence from the model parameters with increasing distance from the mean. Clearly, small changes in the model parameters have more impact on the residuals of "distant" observations. Therefore the residuals ϵ need to be standardized (also known as studentized residuals [62]). For this, observations y and predictions $\hat{y}$ are related by $\hat{y} = Hy$ through the so-called *hat matrix* H. The hat matrix has thus a notion of indicating the influences of each observation y on each of the predicted values $\hat{y}$. From equation 114 it follows that

$$\hat{y} = X\beta = X(X^TX)^{-1}X^Ty \quad \Rightarrow \quad H = X(X^TX)^{-1}X^T \ . \tag{135}$$

Furthermore, the relation between y and the residual ϵ is given by

$$(I - H)y = y - \hat{y} = \epsilon \ . \tag{136}$$

According to equation 135, the hat matrix is symmetric. The so-called *leverages* H_{ii} determine the variance of the i-th residual as $\mathrm{Var}[\epsilon_i] = \sigma^2(1 - H_{ii})$. This can be used to compute the *standardized residual*

$$\tilde{\epsilon_i} = \epsilon_i \cdot \frac{1}{\sigma\sqrt{1 - H_{ii}}} \ . \tag{137}$$

As a result, figures 42 and 43 yield qualitative evidence towards the correctness of the assumptions of the models, a bit more so for the Δh_{bd} than for α_{lt}. Note that only a subset of the data is shown to avoid clutter. The apparent clusters stem from the fact that the subjects were sometimes standing and sometimes sitting, and that the transitions between these states were comparatively short [69]. The plots of the predicted (fitted) values versus the non-standardized residuals support the zero-mean assumption (figures 42a, 43a). For Δh_{bd} one can see that the residual has a relatively constant mean of $0°$ except for the beginning of the domain and values around $45°$. The quality of this assumption is obviously less for α_{lt}, especially so beyond $160°$. All the same, both residual errors seem to follow a normal distribution (figures 42b and 43b). The results are convincing for h_{bd}, whereas outliers are observed for α_{lt}. This is arguably not so much of a problem for the applicability of the model, as most outliers are observed beyond twice the standard deviation and hence in less than 32% of the observations. The notion that the residuals of both response variables each follow a normal distribution is further corroborated by figure 44 which illustrates the residuals in comparison to a normal distribution.

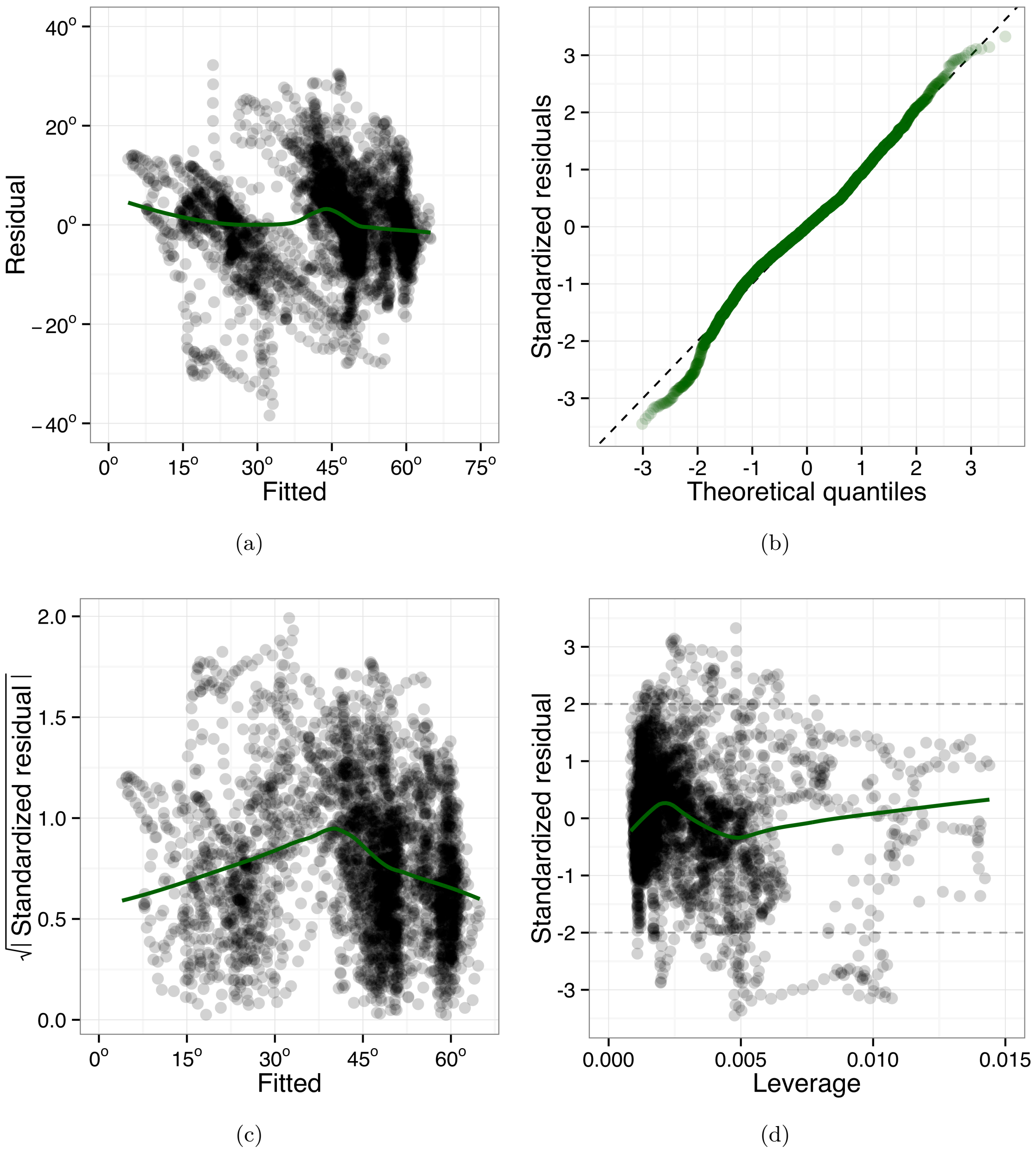
The property of equal variance among the residuals is also called *homoscedasticity.* For linear regression models of the form $y = X\beta + \epsilon$, the homoscedasticity assumption means that variance does not depend on X, i. e. $Var[\epsilon|X] = Var[\epsilon]$, in other words that each observation is equally important for estimating the mean squared error. Figures 42c and 43c illustrate predicted values in relation to the square root of the absolute value of the standardized residuals. Both figures exhibit a systematic trend, suggesting that the model could be improved by a polynomial (e. g. quadratic) term.

Figures 42d and 43d provide a notion of the influence of each observation on the model parameters. This is also interesting in regard of possible outliers, for which Cook's distance may be used as a metric. Cook's distance estimates the influence of a particular observation by determining the effect that removal of this observation would have on the model [62]. The typical thresholds of 0.5 and 1 are outside of the limits of any points though. Also note that the major part of the standardized residuals is well within $\pm 2\sigma$. At the bottom line, in terms of "usefulness", both models show satisfying results with residual standard errors of about $9.7°$ for h_{bd} and $10.2°$ for a_{lt}. Both models are justified in terms of statistical "correctness" according to computational and qualitative analysis.

3.2.4.1 *Discussion*

The main purpose of the proposed system is an estimate of the user's current heading, given only the measurements from consumer-level mobile phone sensors. The user's heading is the direction of the vector orthogonal to the shoulder line and pointing into the direction which the user is facing. Instead of the absolute heading, the underlying linear regression model predicts the difference between the body heading and the device heading which simplifies the process and avoids special treatment of circular variables. An estimate of the absolute heading is therefore easily given by the sum of differential heading and device heading. In this work, the device heading is defined as a function of the measures of orientation along the y- and (negative) z-axes which was done for the following reasons: Generally speaking, the choice of the reference frame is arbitrary. Existing software development kits typically define the phone's heading with respect to its y- *or* z-axis, or switch between those axes whenever a relevant major orientation change is detected. Problems may arise for attitudes close to the gimbal lock, e. g. when the phone is held in a way such that the reference axis is close or even parallel to vector pointing along gravitational force. Also, mobile devices are carried in various locations and/or orientations. According to related work, most people carry their phones in their trousers pockets or in a shoulder bag [150]. The corresponding study does not mention which side of the phone is up, for instance when carried in the front pocket, but it is reported that people are generally apt to protect their phone by carrying it in a position where its front faces their body, so as to protect the phone's screen. So instead of arbitrarily choosing a reference for the heading, the system defines the device heading as a function of both the phones y- and z-axes.

Due to the use of the differential heading h_{bd}, the choice of the reference for the device heading, and the corresponding correction of the phone's measured attitude, the system is invariant to absolute orientation. In principle, the linear regression model can be regarded as an affine transformation, at least when the model only uses the computed yaw, pitch and roll angles as input variables. Others have similarly computed the user's heading in terms of the gravity vector and a PCA of the acceleration signals, projected onto the horizontal plane, in order to determine pedestrian walking direction [179], or by defining fixed transformations for certain locations on the body [142]. In comparison, the proposed system is based on a slightly higher-dimensional model. The addition of a set of temporal features has been shown to reduce the overall residual error which is caused e. g. by motion, particularly so by motion of the leg when the device is carried in the trousers pocket. Nev-



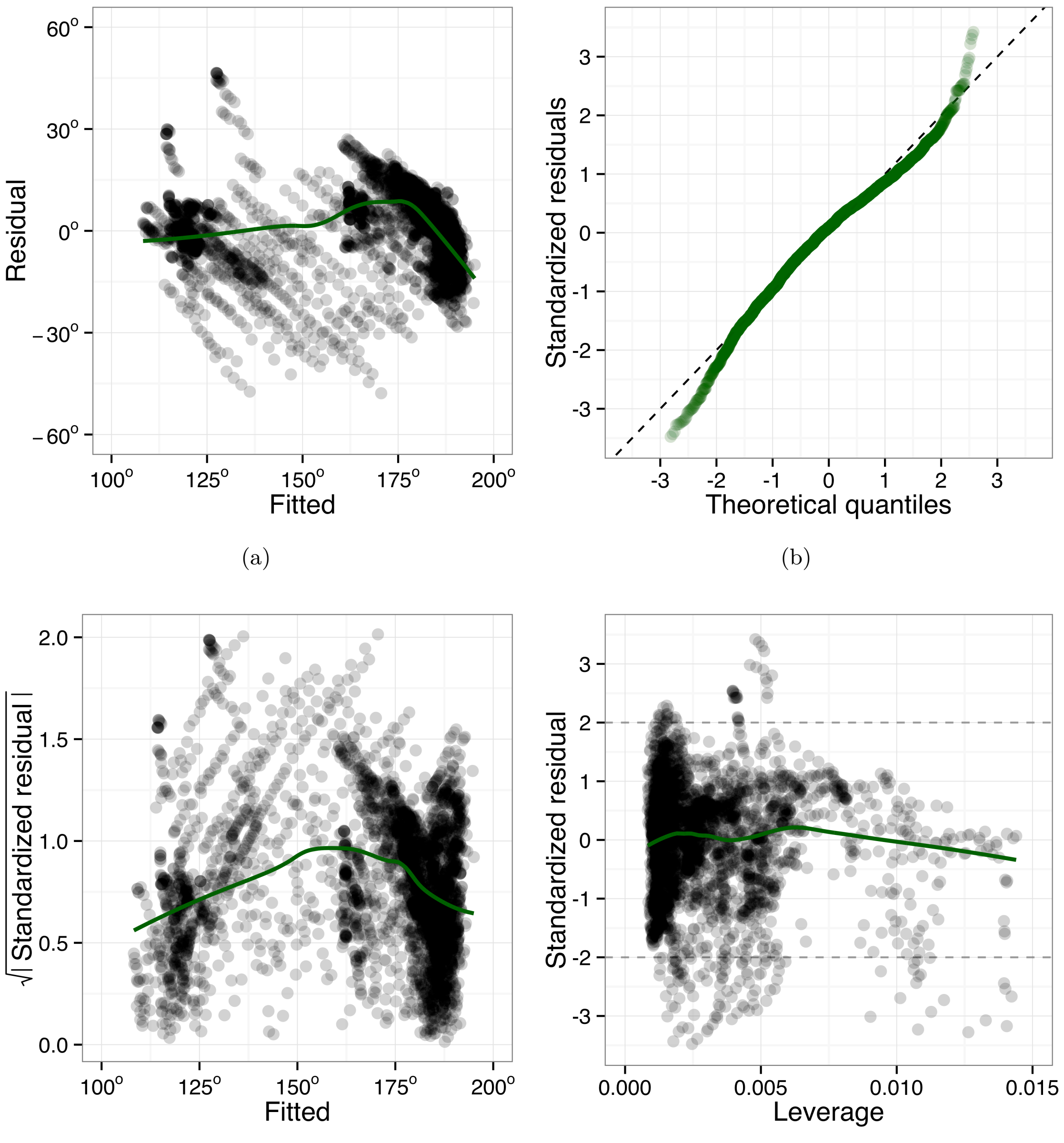
(a)

(b)

(c)

(d)

Figure 42.: Residual analysis for the differential heading, analogous to [69].



(a)

(b)

(c)

(d)

Figure 43.: Residual analysis for the angle between leg and torso, analogous to [69].

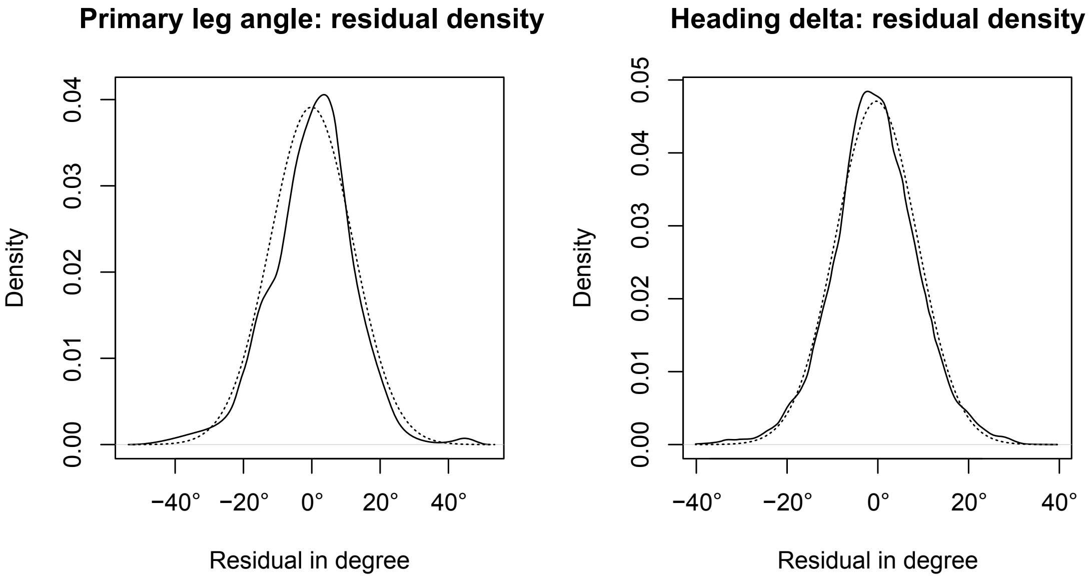


Figure 44.: Distribution of the residuals for the differential heading and the angle between leg and torso. Figure taken from [69].

ertheless, other methods like the PCA-based determination of the walking direction could for instance be used as input to adaptive filters and therefore attribute to the system's overall accuracy.

The model was trained on data from several recordings and persons. Performance will likely increase if personalized models were used instead. It is however unlikely that users have access to the necessary equipment such as the Kinect, or would be willing to undergo a training procedure. Nevertheless it is necessary that more models are created in correspondence to the variability of principle wearing locations and orientations. As [178] have shown, the latter can be accurately determined from patterns in the signals of mobile phone sensors. An application could therefore periodically check for principle changes and adapt by selecting another model accordingly. At last, note that the purpose of developing the proposed system is certainly not outperforming existing systems, although the additional use of temporal features has proven to be beneficial for the overall process. Instead, in the context of this thesis, the system has been developed to prove that algorithmic models for social interaction geometry are not only feasible from a theoretical point of view, but indeed also practicable in "real life" along with present-day consumer hardware. According to the evaluation, the standard residual error of the differential heading is about 9.7°. As a consequence, the computation of the angle $\delta\theta$ between the shoulder lines of two persons is subject to the random errors of two corresponding systems. Residuals are assumed to follow a normal distribution (figures 44, 42). It can be shown that normal distributions

	Predicted				
Actual	$S^{\oplus}$	$S^{\ominus}$	Precision	Recall	F_1-Score
$S^{\oplus}$	279358	88876	81.3%	75.9%	78.5%
$S^{\ominus}$	64321	303913	77.4%	82.5%	79.9%

Table 25.: Confusion matrix after 10-fold stratified cross-validation of a GMM-based classifier with 10 components, assuming Gaussian noise with $\sigma = 13.7°$ on $\delta\theta$ (79.2% accuracy).

are closed under convolution [154]. In other words, the sum Z of two normally distributed random variables

$$X \sim \mathcal{N}(\mu_1, \sigma_1^2) \quad \text{and} \quad Y \sim \mathcal{N}(\mu_2, \sigma_2^2) \tag{138}$$

is itself normally distributed with

$$Z \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2) \; . \tag{139}$$

Since $\mu_1 = \mu_2 = 0$ and $\sigma_1^2 = \sigma_2^2 = 9.7^2$, it follows that the proposed system could predict $\delta\theta$ with a standard deviation of $\sqrt{2} \cdot 9.7 \approx 13.72°$. This is well within limits of applicability for the presented GMM-based model for social interaction geometry, as is easily shown by adding this amount of Gaussian noise onto the test data during 10-fold stratified cross-validation of the model. Indeed, such a model performs very well even in the presence of noise (79.2% vs. 80.3% accuracy, compare tables 25 and 11). Vice versa, these results contribute to the understanding that GMMs are supposed to be a good match for modeling interaction geometry in general, and that the presented model is likely not subject to overfitting in a statistical sense.

3.3 A SYSTEM FOR MEASURING INTERPERSONAL DISTANCE

Section 3.1.2 already introduced several techniques for the estimation and tracking of position and/or measuring distance. Among this related work, Peng et al. [229] have demonstrated a remarkably effective approach for measuring distance for which they used only consumer-level hardware. Based on specifically encoded audio signals, so-called *chirps*, they were able to achieve accuracies of up to centimeter-level. Also recall that their approach does not depend on any special means of synchronization because distance is computed as a function of the time-of-flight of signals between both local/local and local/remote sensors, subject to *a priori* determined systematic delays of the respective hardware (see equation 109). As the signals' time-of-flight is given in terms of audio samples, typical hardware operating at e. g. 44.1 KHz could therefore achieve a theoretical limit of as little as 0.8 cm, assuming 343 ms^{-1} for the speed of sound at $20°C$. Peng et al. also measured the frequency responses of typical consumer-level sensors to chirps from 1 KHz to 20 KHz [229] and found that these are mainly tuned for operations with respect to the vocal spectrum. More specifically, the frequency responses indicated that signals beyond 8 KHz were attenuated too much, which is why they decided to use chirps inside the audible range between 2 KHz and 6 KHz. Despite the benefits of being able to use consumer-level hardware, being independent of special synchronization mechanisms, and achieving centimeter-level accuracy, the use of *audible* signals for measuring distance is without doubt undesirable for SSP scenarios. Ultrasonic methods, on the other hand, will most likely require dedicated hardware, which is also corroborated by the analysis of the frequency responses of mobile devices in [229]. The enormous variety of sensors which are already available in modern mobile phones however suggests near-future applicability of the latter techniques. According to [194], for example, Qualcomm is a hardware manufacturer planning on incorporating ultrasonic sensors in the next generation of consumer-level mobile hardware. The following sections investigate the feasibility of ultrasonic distance measurements in the context of social interaction geometry.

3.3.1 *Prerequisites*

Assuming operations at frequencies of about $40\ \mathrm{KHz}$ and speed of sound of $343\ ms^{-1}$ at $20°C$, the respective wavelengths of $\sim 0.8cm$ yield theoretical sub-centimeter accuracy for ultrasonic range finding sensors. The quality of the results is influenced by several factors such as noise, operating modes, or the precision and granularity of timing devices. Noise may originate from either active or passive sources. For example, other mobile agents which are not yet part of a common distributed network of agents could interfere with an existing, already synchronized, network. Reflections caused by objects or walls constitute a source for passive noise. Walls typically cause diffuse reflections, possibly leading to phase-shift and therefore offsets on the receiver's side. Other than that, timing as well as the granularity of e. g. the system clock plays an important role. The system clock has to allow for operating frequencies higher than those of the sensors. Timing devices must obviously be precise and not prone to systematic or random bias. Systematic errors like the ones caused through computational delays can be regarded as constant and thus be accounted for. However, timing is also important with respect to the operating mode of the sensors. Ultrasonic sensors work in either echo or sender/receiver mode. In echo mode, the sensor functions as a transducer, i. e. it sends and receives its own signal. In sender/receiver mode, one or more dedicated sensors receive the signal burst from yet other sensors. Echo mode thus has the advantage of relying only on a single sensor's internal timing, whereas very accurate and precise synchronization is mandatory for the latter.

Second, recall that speed of sound is proportional to the absolute temperature of a fluid medium, and is independent of density or pressure for ideal gasses. Although air is really not an ideal gas, for the purpose of distance measurements it can be treated as such because the effects of variations in density or pressure are by magnitudes smaller than those of changes in temperature. The speed of sound c can thus be given as a function of temperature ρ $(°C)$ as follows:

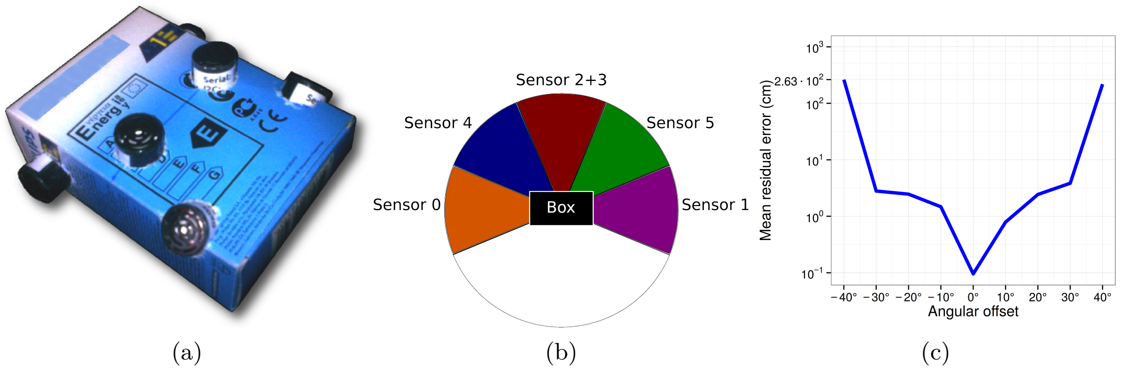
$$c = 20.05\frac{m}{s} \cdot \sqrt{\rho + 273.15K} \tag{140}$$

Reasonable variations in temperature for indoor scenarios may range e. g. from 17°C ($\tilde{3}41\mathrm{ms}^{-1}$) to 23°C ($\tilde{3}45\mathrm{ms}^{-1}$), for which the shift in temperature would yield a maximum error of $8.78 \cdot 10^{-5}$ m at 40 KHz. Taking into account the median of 0.985 m for interpersonal distance δd during social interaction from the present dataset (see section 2.2.5), this shift would cause an offset of merely 1 cm and is thus negligible for applications of models of social interaction geometry. Also note that the location where a mobile device is worn is not important in terms of temperature. Although the device's and thus implicitly the sensor's temperature might change due to the emission of body heat, this has no further influence on the signal's comparatively "long" run between devices through the medium air. In regard of on-body location, it is yet more important to consider a sensor's line of sight and dealing with possible obstructions, which might also constitute a point for further research.

3.3.2 *System configuration*

The proposed system is a proof of concept [202], consisting of up to four small enclosures, each of dimensions $10\mathrm{cm} \times 3.5\mathrm{cm} \times 7\mathrm{cm}$, and each housing an array of 6 ultrasonic sensors. For every sensor box, the sensors are layed out such that they would cover a range of 225°, with two sensors facing the front, two to the side and two at 45° angles (see figure 45). Other configurations are imaginable as a potential result after further analysis of the dataset from section 2.2.5. Since the sensor boxes are supposed to be fastened to a person's belt or hip in experimental setups (near the trousers pocket as the most prominent wearing location [150, 190]), the respective angular offset should be taken into account for all types of configurations. According to the manufacturer, the employed SRF02 ultrasonic sensors feature a beam angle of 55° at -6 dB and are accurate within ± 1 cm from 15 cm up to 5 m, subject to only slight systematic frequency shifts due to temperature [5]. The wide beam could in principle allow for more precise measurements within intersecting areas, for example by averaging the measurements from the respective sensors, and of course provided that the sensors would not operate at the same time, thus avoiding interfering patterns. Control measurements which were performed in advance of any experiments however show that the measurement error in fact grows exponentially beyond 30° (see figure 45c). Each of the sensor boxes is connected to a dedicated linux-based mobile phone via USB. As the sensors are controlled via I2C, each sensor box also contains a USB-to-I2C bridge for communication with the phone.

The system is similar to one of Hazas et al. [138] who had previously used external sensor platforms connected to laptop computers. In addition to ultrasonic sensors, these platforms featured radio-frequency transmitters and receivers, the latter of which were mostly used for synchronization and the communication of small data packets between the devices. Each of their sensor platforms hosted exactly three ultrasonic sensors which were layed out such that they would operate in a mostly two-dimensional layer with sensors at subsequent angles of 90°. Hazas et al. [138] report that they achieved good time synchronization through the use of the RF transmitters. In order to avoid collisions, each laptop kept



(a) (b) (c)

Figure 45.: Placement, coverage and measurement errors with respect to angular offset for the SRF02 sensors (left and middle pictures taken from [202]).

record of all other devices it had seen. After one laptop had sent out its signals, it would then wait for the amount of time that it would need for the known number of other devices to send. Having tested mostly stationary setups, they report accuracies from 6.9 cm to 8.6 cm for distance measurements, and up to 25° for relative orientation in roughly 80% of their measurements, depending on the quality of the line of sight between devices [138]. The key differences to [138] are comprised of a less coarse sensor layout, application in a *mobile* social interaction scenario with much less obstructions through portable devices, and the abstinence from components such as additional RF transmitters. In terms of mobility, one may further note that the laptops in [138] were always firmly placed on top of a table.

3.3.2.1 *Synchronization and sequencing*

As operating the sensors in echo mode was considered impractical for SSP scenarios, sender/receiver (Tx/Rx) mode was used instead. Recall that this however requires precise temporal synchronization. This synchronization would not only involve the exact time where distinct parties would commence sending or receiving, but also preventing internal clocks from shifting apart. Using the aforementioned devices and sensors, it turned out that initial synchronization with a dedicated master device would yield an initial resolution of 50 ms ($\approx 34\ \mathrm{mm}$ at $343\ \mathrm{ms}^{-1}$), but within about 30 s the subsequent shift would go as far as rendering the devices incapable of performing any measurements at all, in spite of using the operating system's high resolution timer in conjunction with real time priority for the process. Wireless broadcasts for continuous synchronization turned out to be useless as well. Random delays of up to $2.5\ \mathrm{ms}$ were observed, thus inducing measurement errors of up to $85\ \mathrm{cm}$. It is assumed that these delays were caused through the implementation of the wireless stack and corresponding parts of the operating system's kernel [202]. Therefore, even though it would mean that the actual implementation of the system would not be independent of external infrastructure anymore, synchronization was eventually achieved by means of externally controlled signals, for which a laptop was connected to each of the mobile phones' audio jacks. The external clock then consisted of an eight-byte audio "impulse" at a frequency of 48 KHz. Although a sound wave would travel ~ $5.72\ cm$ during these 8 samples, the effect is canceled out as it affects both sender and receiver.

Due to the fact that the SRF02 cannot be configured to modulate data payload onto the emitted signal, the sensors and sensor boxes are operated as a token ring. Ranging may occupy a single sensor for up to 66 ms [5], implying an upper bound of 15 measurements per second. In addition to the sensor's own processing, the complete routine that controls the sensor and processes its results takes up to 150 ms. In order to avoid interferences caused by reflections or short-time drift between synchronization points, and taking into account further delays that may be caused by switching agents, or simply through IO operations of the operating system itself, the final polling interval was set to 300 ms. For a network of n agents with k sensors each this means that each sensor will be polled at $n \cdot k \cdot 300\ ms$ intervals. Although the general dynamics of social interaction are typically not considered very high (corroborated by analysis of the datasets in the previous chapters), still, a cycle of this length constitutes a limiting factor in terms of the maximum number of agents. Depending on the application, a (weighted) decision must be made on what type of respective measurement error should be minimized: Is it more important to have accurate readings per person, or should group dynamics be captured as much as possible? The former would imply that at first all sensors of one device should measure, one after another, and only then should the process continue to the sensors of the next device. On the other hand, processing e. g. one frontal sensor of agent A, directly followed by the equivalent sensor of agent B etc. would minimize errors with respect to group dynamics, and minimize the differences between symmetric measurements of δd_{AB} and δd_{BA} at or around one point in time.

3.3.3 *A third dataset*

Yet another series of experiments was conducted for the evaluation of the present proposed system in sequence to those related to the influence of profile and latent parameters described in section 2.4.2. Groups of two, three and four participants were recorded in the same surroundings for about 15 minutes each. In addition to the high-precision infrared tracking system, data from the mobile phones and sensor boxes were recorded as well. For this, the sensor boxes were fastened to the participants' belts or hips as described before. Recall that for the necessary synchronization the devices also had to be connected to a laptop via cables plugged into the phones' audio jacks. The cables were designed to be very thin and long enough so that people could freely move without further obstruction, and layed out such that subjects need not worry to step onto them or otherwise get entangled. As the question may arise at this point, note that neither the sensor boxes nor the cables were present in any of the prior series of experiments.

Each mobile device transmitted a continuous stream of its sensed data via a wireless network. Contrary to the discussed means of synchronization, timing and bandwidth pose no problems for the mere communication of these datastreams. The data from the infrared tracking system were post-processed as described in sections 2.2.2 and 2.4.2, and provide the ground-truth for evaluation of the distance (and possibly orientation) measurements from the ultrasonic sensors. All the same, the data from the ultrasonic sensors were divided into *frames*. For a group of n persons, a single frame consequently consists of exactly $6n$ measurements, and represents an interval of $n \cdot 300\ ms$ starting at time t. For each pair of frame and agent, the median of 4 out of 6 distance measurements was computed, leaving out the minimum and maximum values. This was done for two reasons: First, in order to guard against outliers. And second, in order to satisfy the notion that the chosen layout of sensors would likely yield extrema for those sensors which would either be obscured or otherwise point into an irrevelant direction.

While the main focus is on distance measurements, the positioning of the sensors also allows for a coarse estimation of the direction $\delta\varphi$ where other agents are located. For this, the covered area around a sensor box is divided into nine sectors, each of which is uniquely determined through its adjacent sensors. From left to right, the first sector therefore corresponds to the 45° area around sensor 0, the second sector corresponds to sensors 0 and 1, and so forth (refer to figure 45). For each sector, the readings of the respective sensors are averaged, and the sector with the resulting maximum received signal strength is selected for each pair of agent and frame. Consequently, $\delta\varphi$ is roughly determined as the mean angle of the corresponding sector.

The same manner of dividing the sensed area into nine sectors and evaluating the respective signal strengths also allows for the determination of $\delta\theta$. Let $s_A, s_B \in \{0, 1, \dots, 8\}$ be the indices of the sectors corresponding to the maximum signal strength as perceived by agents A and B. Then $\delta\theta_{AB}, \delta\theta_{BA}$ can be determined as follows [202]:

$$\begin{aligned} \delta\theta_{AB} &= -\text{sgn}(s_A - s_B) \cdot \left(1 - \tfrac{|s_A - s_B|}{8}\right) \cdot \pi \\ \delta\theta_{BA} &= -\text{sgn}(s_B - s_A) \cdot \left(1 - \tfrac{|s_B - s_A|}{8}\right) \cdot \pi \end{aligned} \qquad (141)$$

Note that the lateral and angular offsets of the devices can be neglected because equation (141) describes only angular difference, and position as well as orientation of the sensor boxes were controlled parameters throughout the experiments, i. e. the lateral and angular offsets were the same for both A and B. Without question, a certain systematic error will still remain due to the unavoidable uncertainty when fixating the sensor boxes to the belt or the hips of the subjects. This uncertainty would as well be the case in real-life scenarios, although related work and the prior results have shown that on-body location and orientation of a device can be determined very accurately. Due to the spatial constraints during the recording process (refer to chapter 2), the effect of such a systematic error is however negligible. In case of the present experiment, the infrared tracking system would allow for computing the angular offset as opposed to manual measurements on each participant. For this, that angle of rotation around the yaw axis was determined which would minimize global error in comparison to the infrared-tracking system. The resulting angle corresponds to a counter-clockwise rotation of 47°, and somewhat follows the intuition of

	Residual error	
Variable	Mean	Standard deviation
$\delta\theta$	29.15°	16.93°
$\delta\varphi$	20.28°	31.26°
δd	24.4cm	8.64cm

Table 26.: Mean and standard deviation of the residual for measurements based on ultrasound vs. infrared-tracking

40° ~ 50° for wearing the device on the right hip. As a result, every reading of $\delta\varphi$ and $\delta\theta$ was rotated accordingly.

3.3.4 *Evaluation*

The computation of the values for δd, $\delta\varphi$ and $\delta\theta$, followed by a per-frame comparison of the results with the ground-truth provided by the infrared-tracking system, leads to the results described in table 26. While for each of the three variables the values of the residual's mean and the standard deviation seem surprisingly high, performance still needs to be evaluated with respect to the developed algorithmic model for the discrimination of $S^{\oplus}$ and $S^{\ominus}$ according to social interaction geometry. Recall that the GMM-based models from section 2.2.5 were computed and evaluated based on a significantly larger dataset $\mathcal{D}$, involving groups of two to nine persons over the course of about thirty minutes, and featuring much more group dynamics as the subsequent experiments. The present system was thus evaluated for all sets of variables in $\mathcal{V} = \{\delta d \in \mathcal{X} | \mathcal{X} \in 2^{\{\delta\theta, \delta\varphi, \delta d\}}\}$, i. e. all combinations involving distance measurements. More precisely, cross-validation was performed for each set $\nu \in \mathcal{V}$ on the original dataset $\mathcal{D}$ (refer to section 2.2.5), where for each partition of training- and test-data, Gaussian noise corresponding to table 26 was superimposed onto the respective variables in ν. The results of this evaluation are given in table 27.

Perhaps surprising, in spite of the notable offset and additional noise, for $\nu = \{\delta d\}$ the model performs stable and only slightly less accurate than on the unaltered dataset (refer to table 10). One also notes a slight increase in precision for $S^{\oplus}$, albeit at the cost of recall. For $S^{\ominus}$, on the other hand, recall increases at the cost of precision. For $\nu = \{\delta\varphi, \delta d\}$ overall accuracy is already lower by ~ 10%. This goes along with a huge decrease in recall for $S^{\oplus}$ and precision for $S^{\ominus}$. Yet the result is not unexpected, taking into account the course granularity of $\delta\varphi$ and the exponential increase of per-sensor measurement error beyond angles of 30°, the latter particularly so for true angles "inbetween" adjacent sensors. At last, once $\delta\theta$ comes into play, the performance loss is significant. This is somewhat unexpected when considering the "importance" of each of the variables (refer to section 2.3.5.5). According to the mutual information of $\delta\theta$, $\delta\varphi$ and δd for the discrimination of $S^{\oplus}$ and $S^{\ominus}$, δd is by far the most important, followed by $\delta\varphi$ and only then by $\delta\theta$. On the other hand, mutual information can only serve as an abstract measure and does not

express information about e. g. the distribution of a variable. Therefore, $\delta\varphi$ may still be more important than $\delta\theta$, but the course granularity of both variables as determined by means of ultrasound is not acceptable for $\delta\theta$.
Finally, it can be argued that the mean residuals from table 27 are systematic errors. As such, a second round of cross-validations was performed for which the respective means were disregarded, so that the superimposed Gaussian noise was only determined by the magnitudes of the standard deviations. The results of this second evaluation are given in table 28. This time, the performance for $\nu = \{\delta d\}$ is en par with that of the unaltered dataset. Even once $\delta\varphi$ is taken in addition to δd, accuracy is 10% less, but recall and precision are not affected as much as before.

3.3.5 *Discussion*

Overall, the system's evaluation yields acceptable performance. It has been shown that the main task of the ultrasound system, namely measuring interpersonal distance δd in a social interaction scenario, can be achieved with reasonable accuracy. One may argue that residuals of $24.4 \pm 8.6cm$ are far from reasonable, which may be true in general, but without doubt they constitute a satisfying result for the overall presented algorithmic models for the discrimination of $S^{\oplus}$ and $S^{\ominus}$. They sustain the overall robustness of the GMM-based models and corroborate the choice of these models following the notion that human behaviour is best described in a fluent way as opposed to hard limits. Furthermore, both $\delta\varphi$ and $\delta\theta$ come as a byproduct of the ranging process. In case of $\delta\theta$, the quality of the measurements is insufficient, but the same is not necessarily true for $\delta\varphi$. In any case, these measurements could either serve as a backup, or be combined with those from other systems through e. g. a Kalman filter so as to improve and stabilize the overall process. Cross-validation of the data with super-imposed noise on $\delta\theta$, but not the other variables, shows that merely the combination of ultrasonic-only measurements of both (or all three) variables lead to bad results. Nevertheless, the main focus of the presented system is on distance measurements which could not be provided by other means such as the system from section 3.2, which is why δd cannot be disregarded.

It was argued that the means of the residuals may be seen as systematic errors, and thus be disregarded based on the notion that systematic errors can more or less easily be canceled out. This statement must of course be handled with care. Without doubt, the lateral (dis-)placement of the sensor boxes can be seen as systematic error and the angular offset of each device was corrected according to minimization of the global measurement error. In a real-world scenario such errors with respect to orientation are not always systematic, though. A system such as the one presented in the previous section has of course systematic errors, for example in the underlying model, and those could actually be canceled out. Nevertheless, it is certainly more appropriate to regard residuals of location- and orientation measurements as random errors when they originate from such a system. The same does arguably not hold for distance measurements. Letting aside random errors caused by obstructions, reflections, dynamics, etc., a static bias like the one in the presented system can easily be corrected. One reasonable source for such a bias is e. g. given by the fact that the devices were placed on the right hip. Assuming a constant displacement of the device from the center of the body, the Pythagorean theorem serves to explain a quadratic component of the error, albeit depending on the actual distance. A lateral displacement of 20cm, for example, would result in an error of 12cm for an object at an actual distance of 1.1m.

Another remaining issue is that the presented system is currently restricted to measurements within the forward hemisphere of its carrier. More precisely, the current layout is constrained according to the discussed lateral and angular offsets caused by fastening the device to the user's right hip. In a scenario where two people stand in an L-shaped configuration, the person to the left might simply be unrecognizable for the sensors of the person to the right. The same might hold for people standing in a line or even behind someone else. Hitherto analysis of the datasets in chapter 2 has shown that such interactions occur, even though they are very rare. One could of course at least remotely imagine wearable computing devices such as ultrasonic sensor necklaces, but at present that does not seem convincing. Both problem statements are however relaxed by the fact that at least one of the two devices will be able to see the other. Also recall that models based on a less accurate representation of position and distance, such as the "R2B" model (see figure 25), yield performance in an order of magnitude that mitigates the discussed issues.

Finally, one has to note that the experimental setup was surely not optimal in this case. Although great care was taken in order to avoid obstructions or impose constraints on the participants' behaviour, it is possible that the interactants moved less and were more conscious of the experiment as is. However, the fact that the results of these experiments were transferable to an absolutely independent and much more dynamic scenario/dataset render this critique less crucial. Moreover, when the related works of Peng et al. [229] and Hazas et al. [138] are taken into account, it is most likely that a composition of these three systems would result in a practical system for applications in social interaction geometry.

			$S^{\oplus}$			$S^{\ominus}$		
Variables	Gaussians	Acc	Prec	Rec	F_1-Score	Prec	Rec	F_1-Score
δd	5	77.69%	86.82%	58.92%	70.20%	73.73%	92.79%	82.17%
	10	77.92%	85.55%	60.75%	71.05%	74.38%	91.74%	82.15%
	25	78.60%	84.95%	63.22%	72.49%	75.44%	90.98%	82.48%
	50	77.49%	83.32%	61.95%	71.06%	74.61%	90.01%	81.59%
$\delta\varphi$, δd	5	67.33%	82.44%	33.99%	48.13%	63.92%	94.17%	76.15%
	10	67.91%	82.26%	35.78%	49.86%	64.46%	93.78%	76.40%
	25	68.87%	79.66%	40.59%	53.77%	65.70%	91.65%	76.54%
	50	67.37%	75.70%	39.60%	51.99%	64.85%	89.74%	75.29%
$\delta\theta$, δd	5	53.09%	16.77%	1.27%	2.33%	54.39%	94.82%	69.12%
	10	53.63%	19.00%	1.72%	3.09%	54.67%	95.42%	69.51%
	25	52.34%	24.32%	6.86%	9.48%	54.40%	88.95%	67.34%
	50	47.13%	27.57%	14.52%	18.59%	51.73%	73.39%	60.55%
$\delta\theta$, $\delta\varphi$, δd	5	53.63%	36.84%	7.23%	11.72%	54.92%	91.01%	68.50%
	10	53.87%	34.05%	3.70%	6.54%	54.86%	94.27%	69.35%
	25	50.79%	29.82%	6.96%	10.82%	53.35%	86.08%	65.81%
	50	47.45%	34.30%	20.10%	25.20%	51.94%	69.47%	59.40%

Table 27.: Performance of GMMs with superimposed noise corresponding to ultrasound measurements.

			$S^{\oplus}$			$S^{\ominus}$		
Variables	Gaussians	Acc	Prec	Rec	F_1-Score	Prec	Rec	F_1-Score
δd	5	80.03%	80.09%	73.52%	76.66%	80.00%	85.28%	82.55%
	10	79.95%	79.56%	74.07%	76.72%	80.22%	84.68%	82.39%
	25	80.64%	78.85%	77.33%	78.08%	82.03%	83.30%	82.66%
	50	81.08%	79.03%	78.38%	78.70%	82.71%	83.25%	82.98%
$\delta\varphi$, δd	5	72.05%	77.22%	52.97%	62.84%	69.78%	87.42%	77.61%
	10	71.51%	76.26%	52.46%	62.16%	69.41%	86.85%	77.16%
	25	72.82%	75.33%	58.09%	65.59%	71.51%	84.68%	77.54%
	50	72.31%	74.35%	57.89%	65.09%	71.23%	83.92%	77.05%
$\delta\theta$, δd	5	79.19%	79.90%	71.28%	75.34%	78.72%	85.56%	82.00%
	10	78.68%	79.58%	70.22%	74.61%	78.10%	85.49%	81.63%
	25	79.32%	78.48%	73.90%	76.12%	79.93%	83.68%	81.76%
	50	79.09%	77.73%	74.45%	76.05%	80.11%	82.83%	81.44%
$\delta\theta$, $\delta\varphi$, δd	5	71.42%	76.83%	51.43%	61.62%	69.12%	87.51%	77.23%
	10	70.60%	76.15%	49.63%	60.09%	68.32%	87.48%	76.72%
	25	71.90%	74.67%	55.99%	63.98%	70.51%	84.70%	76.95%
	50	71.40%	73.08%	56.81%	63.92%	70.52%	83.15%	76.31%

Table 28.: Performance of GMMs with superimposed noise corresponding to ultrasound measurements for which the mean systematic error was cancelled out.

4 SENSOR FUSION AND DEDUCTION OF N-ARY SITUATIONS FROM DYADS

4.1 INTRODUCTION AND RELATED WORK

Humans have the ability to assess the presence and quality of social situations very efficiently [166, 226, 336, 271]. In this process, most information is conveyed in a non-verbal manner. Individuals mutually strive to clearly establish or neglect social situations upon entering corresponding scenarios, and once established, they work together to maintain existing situations, for instance when unconsciously compensating for movements of others in ongoing FFSs so as to sustain and/or protect their shared transactional O-space (see section 1.2.2). Nevertheless, all interactants still yield *subjective* perspectives based on which kind and extent of information is available to them, but also dependent on personal context. In turn, their assessments might or might not be known to all or a subset of the other members, and that to varying extent. The other members could then incorporate this information when building their own subjective opinions, weighted by the quality of mutual relation and trust.

Analogously, consider a decentralized Mobile Social Networking (MSN) scenario [353, 249] where each individual is represented by their own mobile agent, e. g. a software agent running on a mobile phone. In such a scenario, an agent could collect measurements from a great variety of physical and logical sensors, leading to its personal *belief* about one or more dependent variables, such as for example the likelihood of the individual represented by the agent being engaged in social interaction with another particular individual, possibly based on algorithmic models of interaction geometry (see section 2.3). Clearly, these personal opinions of an agent are a result of the underlying sensor models, the kind and quality of the involved physical and logical sensors, and the respective uncertainty. Also, not all sensors might be available at all times. Furthermore, different agents might (and likely will) use different sets of sensors. Multiple agents will therefore generally deduce different views on a particular social setting.

Overall, perhaps depending on the type of application, one may presume that agents would greatly benefit from sharing their (raw) data and/or (abstract) subjective opinions. Whereas an in-depth discussion and evaluation of the full consensus process among multiple agents is out of the scope of this work, it will be shown that combining *subjective beliefs* leads to a significant improvement for the question about the participants of social situations. Readers should note that the following is to some extent a recapitulation of the paper “Combining Evidence for Social Situation Detection” by Groh et al. [125], co-authored by the author of this thesis.

4.2 FOUNDATIONS

Classical sensor fusion differentiates between complementary, competitive and cooperative sensors [45, 253]. Whereas fusioning of the sensor outputs is usually not necessary for complementary sensors, it is only naturally to do so for competitive sensors. This is also the classical case in the field of sensor fusion, and as well predominant in decentralized MSN scenarios.

From the perspective of traditional sensor fusion, sensors output a value $\nu(t)$ at any time t within a confidence interval $\epsilon(t)$ with confidence $1 - \alpha$. In a rather trivial scenario where the outputs of two physical sensors which measure the same entity should be fusioned, it is quite common to model the measurement error with a normal distribution. The latter follows from the central limit theorem according to which the sum of multiple random variables converges towards the normal distribution, no matter what the original distributions of those variables might be. Fusioning of the two sensors can therefore be achieved by convolving the respective normal distributions of the two sensors, resulting in yet another normal distribution whose mean corresponds to the expected outcome and whose variance corresponds to the accumulated uncertainty. On a sidenote, this is also the basic principle behind the popular Kalman filter [273, 342]. Fusioning techniques like the one just described can cope with *uncertainty* but lack the ability to model *subjectivity*. Moreover, the given scenario is limited to sets of *homogeneous* sensors. In decentralized MSN it is however likely that distinct agents build their subjective beliefs about the state of a system based on the output of varying sets of *heterogeneous* logical and/or physical sensors which furthermore may or may not be available at all times. For example, one agent might deduce social activity from interaction geometry while another relies on the analysis of turn-taking patterns from audio recordings.

From a much less simplistic perspective, probability theory would allow for modeling sensor networks in terms of Bayesian Networks (BNs), leading to representative and coherent models that are well understood. BNs are easily visualized and therefore also often easy to interpret. Numerous methods exist for filtering, smoothing and/or extrapolating time series of measurements[34, 218] in BNs. Once modeled, a multitude of statistical methods allow for the estimation of the network parameters, such as EM-based methods (refer to section 2.3.1.1). Their nature however requires corresponding sets of training data recorded in either real-life or experimental settings. It is quite clear that the outcome of the final model depends on the quality and quantity of those training data. In regard of applications in SSP, and in particular so for social interaction geometry, there can be no doubt that there will ever only be finite training data which surely cannot represent every possible twist in human behaviour. One may presume though that the modeled aspects of human behaviour are uniform enough in a way that allows for generalization of correspondingly modeled and learned BNs (see sections 2.3.5.6 and 2.4, as well as the results of the evaluation in sections 2.3.5, 2.4.3 and 2.4.3.1). Once the model parameters have been learned, BNs can easily be adopted by mobile agents and would henceforth be essential for their respective view of the world. Yet in spite of their obvious benefits in terms of incorporation and application, e. g.

with respect to the limitations of mobile hardware, BNs seem less suited for the *exchange* of measurements and subjective opinions between agents. Such models would require strong *a priori* knowledge about the participating systems' sets of sensors as well as their precise communication structure, especially so in case of MSN and its presumed heterogeneous infrastructure. It is certainly possible to model a respective BN, yet the required number of model parameters would probably grow enormously, thus – aside from modeling issues – creating an exponential growth in the demand for training data [34, 218].

According to Helton, [141] in [290], there is a dichotomy between *aleatory* and *epistemic* uncertainty. Aleatory uncertainty arises from the fact that a *known* system behaves in a random way, whereas epistemic uncertainty is due to ignorance of the system's exact behaviour. Probability theory usually handles the former through a frequentist approach [218], the disadvantages of which for sensor fusion have been discussed above. Epistemic uncertainty, on the other hand, is modeled with a Bayesian approach, which has the disadvantage that it requires precise knowledge of all system components as well as any possible events, along with a complete set of models for their probabilities. In [290], Sentz gives two examples for further illustration:

- Suppose a system consisting of three components as seen through the eyes of someone who is an expert for only a single component. The expert can make a proposition about the probability p with which this single component might fail. Due to his ignorance of the rest of the system, he might however assume a uniform distribution of failure over the remaining two components.
- Probability theory requires the probabilities for all atomic events to sum up to one. This way the complementary probability is defined for every known event. A reasonable question therefore is whether the same expert would also assume that the *whole* system would *not* fail with probability $1-p$, with p corresponding to only his and $1-p$ to all remaining components.

Sometimes uncertainty cannot be expressed in terms of precise probabilities (as is the case above). *Belief theory* therefore regards probabilities as intervals or sets of atomic events [72, 294, 73, 159, 161, 290]. It extends classical probability theory by the ability to explicitly express *ignorance* [161], which implies three advantages according to [290]:

- Experts are only asked for their precise opinion if they can have one.
- Estimates can be made with respect to multiple events (a set $\mathcal{E}$ of events) without having to resort to giving estimates about particular events (any non-empty subset of $\mathcal{E}$).
- Measures from multiple sources need *not* sum up to one (axiom of additivity). It is possible though, and if that happens, then that would correspond to classical probability theory. If however the sum of the measures were subadditive (less than one), that would imply conflicting information, whereas superadditivity would occur in case of cooperative effects between multiple sources of information.

Based on belief theory, Dempster-Shafer theory (DST) provides a well-known framework for epistemic uncertainty [72, 294, 73] and has "attracted considerable attention" in the field [354]. DST is closely related to probability theory [290], yet due to its roots in belief theory is furthermore capable of expressing lower and upper bounds of probability in terms of *belief* and *disbelief* from the point of view of a single source of information. Furthermore, *Dempster's rule* [72, 294] defines an operation for the fusion of multiple sources of information, as will be discussed next.

4.2.1 *Dempster-Shafer theory*

Assume a state space $\Theta = \{x_1, \dots, x_N\}$, also called *frame of discernment*, along with a given entity A, representing a system which at any time is in one of the mutually exclusive states x_i. From the generalized perspective of an expert (agent), probabilities are not exclusively assigned to atomic states. Instead, a Belief Mass Assignment (BMA) m considers sets of states:

$$m : 2^\Theta \to [0, 1] \tag{142}$$

subject to

$$m(\emptyset) = 0 \qquad \text{and} \qquad \sum_{\theta \in 2^\Theta} m(\theta) = 1\ . \tag{143}$$

Each $m(\theta)$ corresponds to the fraction of the overall evidence that A is in any *one* of the atomic states $x_i \in \theta$. Belief mass is expressed for a particular set θ and does not imply mass assignments for any of its subsets.

Next, let $\theta \in 2^\Theta$ denote a proposition about A being in any one of the states in θ. An agent's *belief* b about θ is defined as

$$b(\theta) = \sum_{\theta' \subseteq \theta} m(\theta') \tag{144}$$

To the contrary, *disbelief* expresses the agent's total belief that A is in none of the states in θ [159]. Hence it sums up all the evidence which speaks against the given proposition

$$d(\theta) = b(\overline{\theta}) = \sum_{\theta' \in 2^\Theta, \theta' \cap \theta = \emptyset} m(\theta')\ , \tag{145}$$

where $\overline{\theta}$ denotes the complement. It follows that the following condition will always hold[72]:

$$b(\theta) \leqslant d(\theta) \tag{146}$$

Plausibility amounts to the total belief in the possibility that θ were true except for the explicit evidence against it, i. e.

$$Pl(\theta) = 1 - d(\theta) = 1 - b(\overline{\theta})\ . \tag{147}$$

The remaining *uncertainty* about θ is lower-bounded by belief and upper-bounded by plausibility. It can therefore be defined as the total belief of superstates or partially overlapping states [159]:

$$u(\theta) = \sum_{\theta' \in 2^\Theta, \theta' \cap \theta \neq \emptyset,\ \theta' \not\subseteq \theta} m(\theta') \tag{148}$$

Assigning all belief mass to Θ yields total uncertainty. It follows that

$$\forall \theta \neq \emptyset :\ b(\theta) + d(\theta) + u(\theta) = 1\ . \tag{149}$$

Using the above definitions, the expected value of the probability of θ is given by

$$p(\theta) = \sum_{\theta' \in 2^\Theta} m(\theta') \frac{\theta \cap \theta'}{|\theta'|}\ , \tag{150}$$

for which $|\theta'|$ is defined as the number of atomic $x_i \in \theta'$ [160]. At this point, equations (149) and (143) lead to interpretability of $b(\theta)$, $d(\theta)$ and $u(\theta)$ as barycentric coordinates. $p(x)$ is then the projection of the point given by the $b(\theta)$, $(d\theta)$ and $u(\theta)$ onto the principal axis of a triangle, connecting disbelief to belief (see figure 46 on page 172). One may further note that whenever lower and upper bound collapse, belief theory reduces to classical probability theory. For further reference, table 29 gives an example of the former definitions for a tri-state system.

4.2.2 *Dempster's rule of combination*

Dempster's rule of combination yields a fusioning mechanism for multiple *independent* sources of information over the same state space Θ, for which the agents each express their own expertise in terms of their subjective BMAs.

Let m_1 and m_2 be the personal BMAs of two independent agents. Their combined BMA m_{12} for a proposition $\theta \in 2^\Theta$ is then defined as

$$m_{12}(\emptyset) = 0 \tag{151}$$

$$m_{12}(\theta) = \frac{\sum_{\upsilon \cap \varphi = \theta} m_1(\upsilon) m_2(\varphi)}{1 - \sum_{\upsilon \cap \varphi = \emptyset} m_1(\upsilon) m_2(\varphi)} \tag{152}$$

The denominator serves not only as a normalization factor, but as a means of ignoring all conflicting information, therefore "attributing any probability mass associated with conflict to the null set", [352] in [290]. As a generalization of Bayes' rule, Dempster's rule behaves like an *AND*-operation by emphasizing the agreement between multiple sensors [290]. Amongst others, Zadeh [354] showed that this normalization factor as a matter of fact causes Dempster's rule to produce unreliable results or to completely fail in case of highly conflicting beliefs [354, 44, 74]. A corresponding example is given by Jøsang [159]: Consider a murder case with three suspects Peter, Paul, and Mary, as well as two witnesses

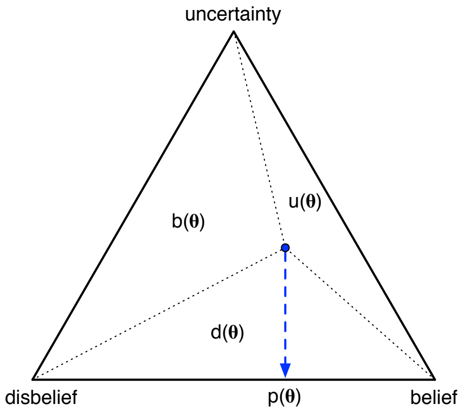


Figure 46.: Belief, disbelief and uncertainty about a proposition θ in terms of barycentric coordinates. The probability $p(\theta)$ is the projection onto the principal axis.

Proposition	Mass	Belief	Disbelief	Plausibility	Uncertainty
$\emptyset$	0.00	0.00	0.00	1.00	0.00
{a}	0.19	0.19	0.49	0.51	0.32
{b}	0.20	0.20	0.61	0.39	0.19
{c}	0.25	0.25	0.48	0.52	0.27
$\{a\} \cup \{b\}$	0.09	0.48	0.25	0.75	0.27
$\{a\} \cup \{c\}$	0.17	0.61	0.20	0.80	0.19
$\{b\} \cup \{c\}$	0.04	0.49	0.19	0.81	0.32
$\{a\} \cup \{b\} \cup \{c\}$	0.06	1.00	0.00	1.00	0.00

Table 29.: Belief theory measures for a system with three atomic states a, b, and c.

	Witness 1	Witness 2	Dempster's rule
Peter	0.99	0.00	0.00
Paul	0.01	0.01	1.00
Mary	0.00	0.99	0.00
Θ	0.00	0.00	0.00

(a) Without uncertainty

	Witness 1	Witness 2	Dempster's rule
Peter	0.98	0.00	0.490
Paul	0.01	0.01	0.015
Mary	0.00	0.98	0.490
Θ	0.01	0.01	0.005

(b) With uncertainty

Table 30.: Outcome of Dempster's rule and SL's consensus operator in a classic example of high conflict (a) vs. the outcome after introducing uncertainty over the whole state space (b). Example taken from [159].

with highly conflicting testimonials. The first witness believes that Peter committed the murder with belief mass 0.99, and that Paul likely did it with belief mass 0.01. The second witness however assigns belief mass 0.99 to Mary, and 0.01 to Paul as well. Application of Dempster's rule eventually eliminates Peter and Mary as suspects and leads to a joint belief mass of 1.0 for the initially highly unlikely Paul (see table 30a). According to Zadeh, it is therefore evident that Dempster's rule "cannot be applied until it is ascertained that the bodies of evidence are not in conflict" [354]. Jøsang and Pope alleviate this by arguing that it may still serve as an approximation in cases of low conflict [162]. In [159], Jøsang has furthermore shown that the introduction of a small amount of uncertainty over the whole state space Θ yields a substantial reduction in conflict and can lead to better results (see table 30b). It is further mentioned that "dogmatic" BMAs which assign zero belief mass to Θ (as opposed to disbelief or uncertainty) are considered hypothetical and "unnatural in practical situations" [159].

4.3 SUBJECTIVE LOGIC

Subjective Logic (SL) is an extension to DST that handles both uncertainty and subjectivity. It serves as a generalization of first-order logic and probability calculus, and it can be shown that it collapses to either one of them whenever the input parameters are chosen accordingly [161]. SL itself is based on the presumption that, in principle, it is never possible to state whether a particular assumption about the world (a system) is true or false with absolute accuracy, in other words that "perceptions about the world are always subjective" [163]. For this, SL regards specific types of beliefs, called *opinions*, and provides a rich set of SL operators [159, 160, 163, 161]. Opinions are expressed using BMAs which assign belief mass only to atomic states $x \in \Theta$ and Θ as a whole, i. e.

$$m(\theta) \neq 0 \;\Rightarrow\; \theta = \Theta \vee |\theta| = 1 \, , \tag{153}$$

known as Dirichlet Belief Mass Assignments (DBMAs). Consequently, for DBMAs belief b is always equal to the basic mass assignment m for all atomic states x_i. Defining uncertainty only for the x_i and Θ makes it explicit that no one will ever be able to state assumptions about the world (system) with absolute accuracy. This presumption also leads to another representation of the expected probability for a given proposition. From equation 150 one can see that the remaining belief mass from (partially) overlapping states is distributed in equal kind for $\theta \subseteq \Theta$ of the same cardinality $|\theta|$, in particular so for atomic states $x_i \in \Theta$. In case of DBMAs, the only such overlapping state is Θ itself. It follows that $m(\Theta)$ must be distributed *uniformly* among all atomic states x_i for arbitrarily large frames of discernment. This insight can be captured as a function

$$a : \Theta \to [0,1] \quad \text{subject to} \quad a(\emptyset) = 0 \qquad \text{and} \qquad \sum_{x_i \in \Theta} a(x_i) = 1 \; . \tag{154}$$

This function is called the *base rate* function and naturally represents the *a priori* probability for each of the x_i to be true, i. e. their probability in absence of any evidence [163]. According to Jøsang, the default base rate function corresponds to a uniform distribution, but there is no requirement for that. Different opinions over the same state space may share the same base rate function, except for e. g. situations where distinct analyses of the same Θ need to be modeled for different persons [161].
Another benefit that follows from the fact that DBMAs distribute belief mass only among the x_i and Θ is that uncertainty, as previously defined in equation (148), can now be simplified to a single scalar value

$$u = m(\Theta) \; , \tag{155}$$

so that the expected probability from equation (150) can be expressed as

$$p(x_i) = b(x_i) + u \cdot a(x_i) \; , \tag{156}$$

i. e. the posterior probability of A being in state x_i. It can be shown that $p(x_i)$ is a valid probability density function, considering equations 143 and 153 due to which $\sum_i b(x_i) + b(\Theta) = 1$ for all DBMAs.

At last, the prior prerequisites lead to the definition of an *opinion* over a given state space Θ as as three-tuple

$$\omega_\Theta = (\mathbf{b}, u, \mathbf{a}) \; . \tag{157}$$

The vectors $\mathbf{b}$ and $\mathbf{a}$ denote DBMA and base rate, and the scalar u defines uncertainty. Opinions over arbitrarily large state spaces are called *multinomial*, those over binary state spaces are referred to as *binomial*. For multinomial opinions, $p(x)$ follows a Dirichlet distribution, whereas it follows a Beta distribution for binomial opinions [161]. A mapping from subjective opinions over binary state spaces to Beta distributions is given in [163]. The preferred notation for binomial opinions as a four-tuple

$$\omega_{\Theta_{binary}} = (b(x), b(\bar{x}), u, a(x)) = (b, d, u, a) \tag{158}$$

for the binary state space $\Theta = \{x, \bar{x}\}$. Although including the disbelief is clearly redundant (refer to equation (149)), it allows for a more compact operator notation. Note that binary state spaces are also referred to as *focussed frames of discernment.*

4.3.1 *Fusion operators*

SL provides a rich set of operators for opinions as input and output parameters. While most operators are related to well-known operators from logic and probability calculus, some exist only in the context of SL. A comprehensive listing of SL operators can be found in [161]. The fact that SL is compatible with both logic and probability calculus can be shown by considering frames of discernment such that e. g. the input and output parameters correspond to binary TRUE or FALSE. In such cases the SL operators will produce the same results as propositional logic does. The equivalent holds for probabilities [159, 160, 163, 161]. For applications in SSP, the most interesting operations are implemented by the *cumulative* and *averaging* fusion operators. Cumulative fusion is the case whenever agents observe the same process at different times. This also means that they are considered to be independent sources of information, which is also an important presumption in Dempster's rule (see section 4.2.2). Averaging fusion, on the other hand, is concerned with agents observing the same process simultaneously or within at least partially overlapping time frames. For (partially) dependent observations, a *hybrid* fusion operator can be defined [163].

CUMULATIVE FUSION OPERATOR Let ω^A_Θ and ω^B_Θ be the two opinions of sources A and B over the same multinomial state space Θ. The *cumulative fusion operator* $\omega^A \oplus \omega^B$ is defined as

- $u^A \neq 0 \vee u^B \neq 0$:

$$\omega^A \oplus \omega^B = \omega^{A\diamond B} = \begin{cases} b^{A\diamond B}(x_i) & = \frac{b^A(x_i)u^B + b^B(x_i)u^A}{u^A+u^B-u^Au^B} \quad \forall x_i \in \Theta \\ u^{A\diamond B} & = \frac{u^Au^B}{u^A+u^B-u^Au^B} \end{cases} \tag{159}$$

- $u^A = 0 \wedge u^B = 0$:

$$\omega^A \oplus \omega^B = \omega^{A\diamond B} = \begin{cases} b^{A\diamond B}(x_i) & = \gamma^A b^A(x_i) + \gamma^B b^B(x_i) \\ u^{A\diamond B} & = 0 \end{cases} \tag{160}$$

where $\gamma^A = \lim_{u^A\to 0, u^B \to 0} \frac{u^B}{u^A+u^B}$ and $\gamma^B = \lim_{u^A\to 0, u^B \to 0} \frac{u^A}{u^A+u^B}$.

This definition follows the notion that the fusioned opinion $\omega^{A\diamond B}$ should equal that opinion which yet another agent C would have after monitoring all events of the process during both time frames. The operator can be understood as a generalization of SL's consensus operator which is defined over binary state spaces. [160] illustrates the use of the consensus operator in spite of ternary state spaces, such as e. g. in the example of the three murder suspects. It should be pointed out that the second case corresponds to a complete lack of uncertainty (as in probability calculus). The result resembles a weighted average of probabilities, and is equal to the result of combining two measurements under the assumption of normally distributed measurement errors (see section 4.2).

AVERAGING FUSION OPERATOR Let ω_Θ^A and ω_Θ^B be the two opinions of sources A and B over the same multinomial state space Θ. The *averaging fusion operator* $\omega^A \underline{\oplus} \omega^B$ is defined as

- $u^A \neq 0$ or $u^B \neq 0$:

$$\omega^A \underline{\oplus}\, \omega^B = \omega^{A \underline{\diamond} B} = \begin{cases} b^{A \underline{\diamond} B} & = \frac{b^A(x_j)u^B + b^B(x_j)u^A}{u^A+u^B} \quad \forall x_i \in \Theta \\ u^{A \underline{\diamond} B} & = \frac{2u^A u^B}{u^A+u^B} \end{cases} \tag{161}$$

- $u^A = 0$ or $u^B = 0$:

$$\omega^A \underline{\oplus}\, \omega^B = \omega^{A \underline{\diamond} B} = \begin{cases} b^{A \underline{\diamond} B} & = \gamma^A b^A(x_j) + \gamma^B b^B(x_j) \\ u^{A \underline{\diamond} B} & = 0 \end{cases} \tag{162}$$

where $\gamma^A = \lim_{u^A \to 0, u^B \to 0} \frac{u^B}{u^A+u^B}$ and $\gamma^B = \lim_{u^A \to 0, u^B \to 0} \frac{u^A}{u^A+u^B}$.

4.3.1.1 *Properties*

Both the cumulative and the averaging fusion operator are commutative, associative and non-idempotent [161]. This is an important property because order should not matter when combining beliefs, a fact not not necessarily true for other published combinators [160]. It can furthermore be shown that both operators satisfy equation 149. For comparison with DST, table 31 illustrates the application of the cumulative fusion operator over the murder-suspect state space. The improvements through introduction of uncertainty over the whole system have been demonstrated in tables 30a and 30b. It comes to no surprise that the results of Dempster's rule in 30b are quite similar to those of SL consensus. The difference is subtle, but this is not always the case [160]. To see this, assume a belief mass over a binary state space $\Theta = \{x, \bar{x}\}$ which is distributed such that $m(\{x\}) = 0.9$ and $m(\Theta) = 0.1$. In particular, no belief mass is associated with $\bar{x}$. This setup can be interpreted as an expert stating their opinion about x being true and overall uncertainty about Θ as a whole, yet being reluctant to state their expertise about x being false. In this example, the results differ vastly between both frameworks. Although they are quite similar for x and $\bar{x}$, they greatly differ for Θ. Dempster's rule amplifies the belief in x much more than SL's consensus operator. To the contrary, SL also takes into account the overall uncertainty about Θ, which comes much closer to an intuitive way of thinking.

	Witness 1	Witness 2	Dempster's rule	SL consensus
Peter	0.98	0.00	0.490	0.492
Paul	0.01	0.01	0.015	0.010
Mary	0.00	0.98	0.490	0.492
Θ	0.01	0.01	0.005	0.005

Table 31.: Outcome of Dempster's rule and SL's consensus operator in a classic example of high conflict (a) vs. the outcome after introducing uncertainty over the whole state space (b). Example taken from [159].

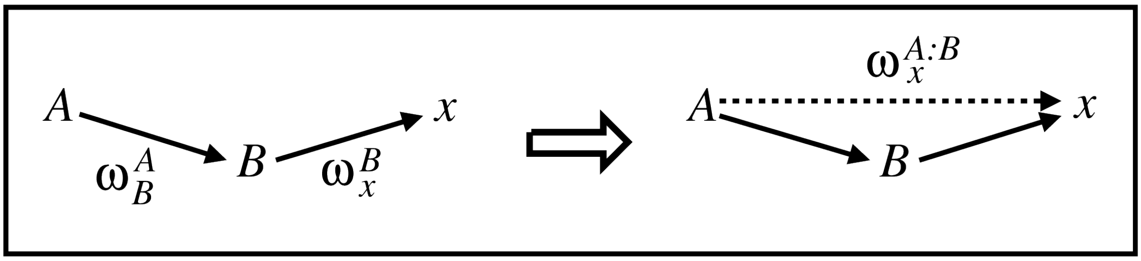


Figure 47.: Transitivity of trust through the discounting operator. Figure taken from [163].

4.3.2 *Trust modeling*

So far, it has been discussed how SL is used for the explicit treatment of uncertainty through subjective opinions. Apart from modeling uncertainty and belief ownership though, another useful property is SL's ability to model trust. The latter proves to be especially useful in the context of a MSN scenario such as the detection of socially interacting groups based on the subjective opinions of multiple independent agents. For the purpose of trust modeling, the idea is to interpret trust as "belief in reliablity" [163], and consequently enabling SL – as a calculus for belief – to be used for reasoning about trust. Practically speaking, *trust transitivity* is achieved by combining individual opinions about trust with other opinions about particular elements of a given frame of discernment. For this, trust in other agents is modeled over a binary state space $\Theta = \{B, \bar{B}\}$, so that ω^A_B corresponds to the belief of A in whether B is reliable (or not, as in $\bar{B}$). Linking w^A_B with the corresponding opinion w^B_x of B about x by using an appropriate operator would then lead to the *transitive opinion* $w^{A:B}_x$ of A about x, as illustrated in figure 47.

Careful consideration of the available operators is mandatory because the opinions of multiple agents might depend on each other without the agents being aware of that. For example, several unrelated newspaper authors might have listened to the same secret source of information. In cases were mutual dependencies are likely, operators must therefore be chosen accordingly as already indicated in section 4.3.1. Failure to compensate for dependent sources might otherwise result in some opinions being emphasized beyond reason. However, according to Jøsang et al. there is no single flawless operator for trust transitivity due to the insight that "trust transitivity is a psychological phenomenon that cannot be objectively observed" [163]. Following these considerations, Jøsang et al. present the following two operators for trust modeling [163].

UNCERTAINTY FAVOURING DISCOUNTING Let A and B two agents and Θ the frame of discernment. The idea of the *uncertainty favouring discounting operator* is that *uncertainty* about a proposition $x \in \Theta$ is favoured based on the principle *disbelief* of A in B.

In this regard, A assumes that B has either no knowledge of x or simply ignores the true value of x [163]. Therefore A will ignore x as well.

$$\omega_B^A \otimes \omega_x^B = \omega_{A:B}^x = \begin{cases} b_x^{A:B} & = b_B^A b_x^B \\ d_x^{A:B} & = b_B^A d_x^B \\ u_x^{A:B} & = d_B^A + u_B^A + b_B^A u_x^B \\ a_x^{A:B} & = a_x^B \end{cases} \tag{163}$$

OPPOSITE BELIEF FAVOURING DISCOUNTING Let A and B two agents and Θ the frame of discernment. The *opposite belief favouring discounting operator* serves the case where A believes that B is prone to constantly telling the *opposite* of the true value of x. The operator therefore combines the disbelief of A in B's belief in x with the belief of A in B's disbelief in x.

$$\omega_B^A \otimes \omega_x^B = \omega_{A:B}^x = \begin{cases} b_x^{A:B} & = b_B^A b_x^B + d_B^A d_x^B \\ d_x^{A:B} & = b_B^A d_x^B + d_B^A b_x^B \\ u_x^{A:B} & = u_B^A + u_x^B (b_B^A + d_B^A) \\ a_x^{A:B} & = a_x^B \end{cases} \tag{164}$$

4.3.2.1 *Properties and Use-Cases*

Readers should be aware that Jøsang et al. have defined both operators in terms of the same mathematical symbol. Also note that both operators are associative but not commutative, which otherwise would be nonsensical since trust is inherently directed based on its human nature.

Bamberger et al. [23] use SL to model trust and uncertainty in a car-to-car scenario. The scenario assumes that cars meet multiple times and therefore build a "social structure". Trust is deliberately expressed on an individual basis and there is no such thing as a general reputation. As an example, consider a DBMA over the binary state space $\Theta = \{x, \bar{x}\}$, where x and $\bar{x}$ represent the presence or absence of a new traffic sign at a given location. Individual cars build their opinion about x (or $\bar{x}$) by taking into account all available opinions of the remaining cars. Once enough information has been collected over time, the true value of the accordingly fusioned opinion can be assumed as fairly certain and thus be used to *develop trust* in each of the other cars. Each individual trust relation will then depend on the quality of the other cars' assessments of the proposition in question. For this, given two cars A and B, the trust model consists of the three components competence, predictability, and recommendation [23]:

Competence: A's opinion about B's competence is determined by the mean error of B's opinions after consideration of all available evidence so far, and is denoted as $\omega_{comp,B}^A$.

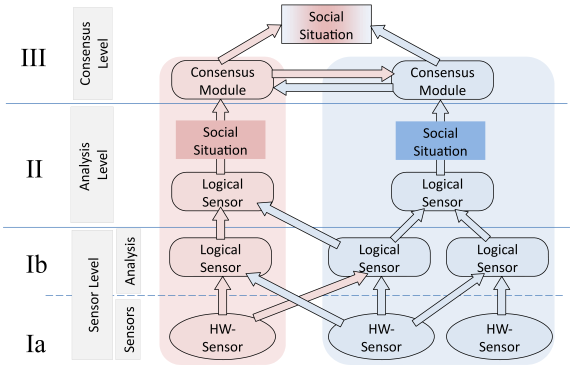


Figure 48.: Topology of the proposed sensor model. Image reproduced from [125].

Competence is therefore inverse to the magnitude of the accumulated mean error. Older evidence is weighted lower by using a time-dependent factor.

Predictability: A's opinion $\omega^{A}_{pred,B}$ about the predictability of B reflects the ability of A to make correct decisions based on B's opinion, considering A's trust in B.

Recommendation: A's opinion $\omega^{A}_{rec,B}$ about B is published as a recommendation upon contact with other cars to help them build or sustain their own *local* reputation of B. Recall that there is no general reputation maintained by any kind of central unit.

4.4 SENSOR MODEL

For applications in a decentralized MSN scenario, the work at hand proposes a sensor model comprised of physical and logical sensors arrayed at different levels of abstraction. *Physical sensors* provide raw measurements. They represent the kind of sensors which are directly employed on modern mobile hardware, such as e. g. microphones, accelerometers, compasses, gyroscopes, etc. The sensor model assumes that any agent A_i is equipped with a number $1 \leqslant m \leqslant M$ of these hardware sensors, denoted as H^i_m. *Logical sensors* abstract over any other type of sensor, i. e. both physical and logical. The input for logical sensors is furthermore not restricted to the output of one agent's individual device (local). Instead, they can freely gather and exchange data from and with other devices (remote). These sensors will be denoted as L^i_n, where $1 \leqslant n \leqslant N$ for an agent with N logical sensors. Figure 48 illustrates the topology of the proposed sensor model. On top of the base layer Ia of hardware sensors, layer Ib represents a set of logical sensors, best described in terms of three main categories: First, those logical sensors which directly interpret or combine measurements from one or more *local physical* sensors. For example, consider a sensor $L^i_{\theta_1}$ which integrates the smoothed measurements from a mobile device's accelerometers, gyroscopes and magnetic compasses so as to determine its absolute heading (for an in-depth discussion refer to chapter 3). Second, those that combine several *local logical* sensors, such as a sensor $L^i_{\theta_2}$ which yields the upper body orientation of the person wearing the device with respect to an ENU reference frame. That sensor would e. g. combine the former $L^i_{\theta_1}$ with another logical sensor L^i_b, the latter of which reflecting precise on-body location and orientation of the device (corresponding sensors were discussed in section 3.2). Third, those sensors that relate the outputs from both *local and remote physical* sensors. An example for this type would be a sensor $L^i_{\delta d;i,j}$ which interprets the readings from two ultra-sound sensors in order to determine the interpersonal distance δd_{ij} between agents A_i and A_j (refer to section 3.3 for an in-depth discussion). Further note that none of the sensors on the first level take into account the outputs of any remote sensors from the same level. This design was chosen to emphasize the notion that those sensors which abstract over others typically serve a more analytical purpose (layer II). A sensor on the second level would therefore e. g. combine the outputs of several local and remote logical sensors, such as interpersonal distance $L^i_{\delta d;i,j}$, upper-body orientations $L^i_{\theta_2}$ and $L^j_{\theta_2}$, and relative positions $L^i_{\delta\phi;i,j}$ and $L^j_{\delta\phi;j,i}$, in an effort to determine whether A_i and A_j interact. For this, a sensor $L^i_{GEO;i,j}$ would e. g. employ the model for social interaction geometry from chapter 2.

In the context of SSP, the proposed model assumes that every agent possesses one or more level II sensors $L^i_{n;i,j}$, each of which yielding the probability with which A_i and A_j do ($p^{\oplus}$) or do not ($p^{\ominus}$) interact. The sensors L^i_n are based on distinct (independent) sources of information. Aside from the example given in the previous paragraph, this could be as trivial as inferring one such pair of probabilities from Bluetooth encounters within a given time frame, however unreliable that may be, or via the analysis of turn-taking patterns from audio recordings. The latter technique was successfully demonstrated by Groh et al. in [126]. Yet another technique which focusses on the correlation of low-level Mel Frequency Cepstral Coefficients (MFCCs) from level Ib audio sensors was developed in the proceedings of [234]. The goal of this approach is to come up with a set of social situations as a result of incorporating the Social Network (SN) inside the social sphere [119] around an agent. In principle, the social sphere includes any other close-by agents, for instance as determined via Bluetooth, Wifi networking, ultrasound ranging, or other appropriate means of near-field communication.

The model assumes that level II logical sensors are eventually combined by one top-most logical sensor per agent. The output of this sensor corresponds to the agent's belief about its own social situation, as well as those of the perceived agents. For this, the top-most logical sensors $L^i_{n;i,j}$ communicate their outputs as a function of time t in the form of subjective opinions $\omega^i_{n;i,j}(t) = (b, d, u, a)$. Belief, disbelief and uncertainty directly relate to the probabilities $(p^{\oplus}, p^{\ominus})$, i. e. for or against agents A_i and A_j participating in the same social situation. For this, the base rate a represents the *a priori* probability for the pair of A_i and A_j. From a naïve point of view, it could be chosen as the default base rate, i. e. according to a uniform distribution [161]. Instead, a better intuition of a is given by the number x of past encounters between A_i and A_j in relation to the total number y of encounters that A_i has experienced within a predetermined time frame $[t-\tau, t]$. The base rate can then be expressed conveniently as $a = \frac{x+1}{y+2}$. Following the discussion about the relevance of group size in section 2.4, future attempts might further consider incorporating respective priors, realized e. g. in terms of a corresponding logical sensor.

4.5 SOCIAL SITUATION ESTIMATES

The proposed model uses SL fusion operators to combine outputs from the logical sensors. Recall that these outputs are in the form of subjective opinions. It can be assumed that these opinions are partially dependent due to the fact that some sensors, such as $L^i_{\delta d;i,j}$ providing ultra-sound based distance measurements between A_i and A_j, rely on mutually dependent measurements from both local and remote sensors in order to produce a result. In addition to this it is also important to model two different aspects of trust. First, trust in sensors L^i_n is modeled as a function over time $\omega^i_{L^i_n}(t)$. Second, following the discussion in section 4.3.1, A_i's and A_j's logical sensors are combined through averaging fusion and the uncertainty favouring discounting operators from section 4.3.1:

$$\omega^i_{\{i,j\}}(t) = \left[\underline{\bigoplus}_{n \in N_i} \left(\omega^i_{L^i_{n;i,j}}(t) \otimes \omega^i_{n;i,j}\right)\right] \underline{\bigoplus} \left[\underline{\bigoplus}_{n \in N_j} \left(\omega^j_{L^j_{n;j,i}}(t) \otimes \omega^j_{n;j,i}\right)\right] \quad (165)$$

Equation 165 shows that a) the opinions of A_i about its own sensors are weighted by its respective trust through $\otimes$, and b) that the output of A_j's sensors are weighted analogously, in particular also according to A_i's individual trust in each of these now remote sensors. The use of $\oplus$ for fusioning reflects the notion of partially dependent opinions. Arguably, modeling trust separately for every foreign sensor $L^j_{n'}$, seems cumbersome and unfeasible. As an alternative, trust could be consolidated into a single binary opinion $\omega^i_j(t) = (b, d, u, a)(t)$, representing A_i's principle belief in the realiability of A_j. Belief, disbelief and uncertainty parameters of $\omega^i_j(t)$ can be determined analogously to Bamberger et al. [23], i. e. by evaluating statistics on A_j's recent opinions versus a general consensus achieved among multiple agents. For the base rate, an unbiased estimator is given by the default base rate $a = 0.5$ [163].

In order to get the complete picture of the set of social situations within its social sphere, A_i requests the opinions $\omega^k_{\{k,l\}}(t)$ of all perceived agents A_k about their presumed social situations with other agents A_l and $i \neq l$. Each of these opinions is subsequently weighted with the trust $\omega^i_k(t)$ of A_i in A_k. Once all relevant data have been acquired, A_i will then use all trust-weighted opinions to build its own current assessment of $SS_k(t) = (P_k, T_k, X_k, K_k)$ about the social situations for all A_k at time t.

The sets of persons P_k are estimated by considering all $\omega^i_{\{m,n\}}$. For this, a weighted graph $G(t) = (V, E, w)$ is constructed whose vertices V denote the agents, E corresponds to the edges between agents, and $w : E \rightarrow \mathbb{R}$ assigns weights to these edges as a result of evaluating the expected probabilities of the $\omega^i_{\{m,n\}}$. This graph consequently expresses a probabilistic view of the situational SN as perceived by A_i, therefore in particular also including its subjective view on its own SS_i. The graph is then clustered using an algorithm for the detection of non-overlapping components [102, 220], following the definition of non-overlapping social situations as a consequence of full mutual awareness of all participants (refer to chapter 1). The algorithmically determined $SS_k(t)$ are eventually broadcast to the nearby agents. From this point onwards, every agent therefore has a complete (subjective) view of the surrounding situations.

Agents can now deliberate on a consensus which could then be used for several purposes:

For example, agents could improve and/or assess the quality of their own opinions. Agents could furthermore adapt their trust into their own sensors, or likewise for building trust in other agents. Last but not least, each A_i could propagate their trust to enable others to build local reputations for all agents that A_i had contact with (similar to [23]).

4.5.1 *A Protocol for Finding Consensus*

In order to answer the question regarding how agents may eventually achieve a consensual view, the problem is broken down into two phases. Phase one is concerned with finding a suitable set of agents to deliberate with, whereas phase two is about achieving consensus. A full specification of the algorithm and communication protocols can be found in [100]. The proposed protocol for the first part is as follows: It is presumed that agents have means of precisely and accurately synchronizing time. A consensus request will be triggered periodically at isochronous intervals of length τ, i. e. a consensual view should be found for the time frame $[t-\tau, t]$, where t denotes the current time. With respect to a reasonable workload and above all to avoid noisy results, the interval τ should be chosen as a good compromise between avoiding the former yet still allowing for capturing of dynamic social events. It therefore requires to be chosen heuristically (e. g. $\tau = 5s$) and may depend on overall social context. Following the previously described fusion and exchange of opinions, each agent A_i performs smoothing of its subjective estimation $SS_i = (P_i, T_i, X_i, K_i)$ based on its recent estimations. The protocol then defines the following steps:

1. Each agent A_i pushes SS_i to all $A_j \in P_i$, and requests their SS_j in return. In doing so, A_i can be confident that it will receive an exhaustive view of all candidate social situations.

2. Let $S_{req} = \bigcup_k \{SS_k\}$ be the set of social situations received upon request, and let $S_{push} = \bigcup_l \{SS_l\}$ be the set of social situations received as a result of other agents pushing their most recent estimations to A_i. In other words, S_{push} refers to the accumulation of those SS_l that were pushed from A_l to A_i due to A_l's individual execution of step 1. S_{push} will therefore typically be a superset of S_{req} (unless one or more agents $\in P_i$ suddenly cease to operate) and may include additional SS_l from those A_l that A_i is not yet aware of but who themselves consider $A_i \in P_l$. Then $S_{SocSph} = S_{req} \cup S_{push} \cup \{SS_i\}$ denotes the set of candidate social situations inside A_i's social sphere. A_i consequently pushes a *consensus initiation request* to all $A_j \in S_{SocSph}$.

3. All agents A_i mutually accept or decline these requests following a decision function df. The idea behind df is to serve as a distance measure for the social situations $SS \in S_{SocSph}$, effectively partitioning this set into "accepted" and "declined" estimates, ultimately accepting only those estimates that include the optimal candidate situations involving A_i. For the current application, df is proposed as weighted Manhattan distance

$$df(S_i, S_j) = w_P d(P_i, P_j) + w_T d(T_i, T_j) + w_X d(X_i, X_j) + w_K d(K_i, K_j)\ . \tag{166}$$

The choice of the weights w for P, T, X and K is left to the agents and can be adapted to the requirements of the application. Note that the distance metrics vary for each of the variables. For the set P of persons, a component-based Jaccard distance

$$d(P_i, P_j) = \frac{|P_i \cap P_j|}{|P_i \cup P_j|} \tag{167}$$

is suggested, but could as well be replaced by other suitable metrics. Furthermore, since both the temporal reference T and the spatial reference X can be seen as a projection of a higher-dimensional spatio-temporal entity $\tilde{X} \in \mathbb{R}^4$, the suggested distance measure exploits this fact through a relating density between $\tilde{X}_i$ and $\tilde{X}_j$, namely

$$d(\tilde{X}_i, \tilde{X}_j) = \frac{\int \min(\rho_i(x), \rho_j(x)) d^4x}{\int \max(\rho_i(x), \rho_j(x)) d^4x}, \tag{168}$$

for which the $\rho(x)$ denotes spatio-temporal densities. A discrete approximation of the $\rho(x)$ could e. g. be achieved by means of location measurements at precise time intervals of each involved agent.

4. From these data, each agent infers a star-shaped SN graph, induced through mutual agreement.

5. The set of agents which will then enter the consensus phase is finally determined by application of the *Modularity* algorithm [148].

After determination of the sets of agents who will negotiate with each other, the second step is concerned with the acquisition of a consensual view on the social situations. This can e. g. be achieved through yet another application of SL *averaging* fusion. If, in spite of the distance metrics which were applied during initiation of the consensus phase, the subjective estimates within a group of deliberating agents are way too inhomogeneous, then either the alternative approaches of Rosenschein [274] or *cumulative* fusion could be chosen instead. Recall that the latter requires independent sources of information. However unlikely in a SSP scenario, this could still be a case for estimates which do not rely on mutual measurements from either agent's remote sensors.

4.6 EVALUATION

The proposed model was evaluated on the primary interaction geometry dataset from section 2.2, thus allowing for comparison of the results achieved through sensor fusion with those from the previous evaluations (refer to 2.3.5). The concrete sensor model based on the primary dataset is as follows: Layer Ib logical sensors $L^i_{GEO;i,j}$ continuously assess whether two agents A_i and A_j are members of the same social situation. These logical sensors are based on pairwise interaction geometry. More precisely, they rely on the outputs of both local and remote layer Ia and Ib sensors from which $\delta\theta$, $\delta\varphi$, and δd can be observed. In order to decide for $S^{\oplus}$ or $S^{\ominus}$, the $L^i_{GEO;i,j}$ then employ the model for interaction geometry as described in section 2.3. Next, recall that synchronized audio streams were recorded for each subject during the experiment (see 2.2). These audio streams are captured by additional layer Ib sensors L^i_{MFCC} yielding the low-level MFCCs [287]. The MFCCs are computed at a frequency of 2 Hz for a sliding window over the interval $[t-60s, t]$. The outputs of the L^i_{MFCC} are then processed by layer II sensors $L^i_{AUDIO;i,j}$ which focus on the correlation of the pairwise $\left(L^i_{MFCC}, L^j_{MFCC}\right)$. The $L^i_{AUDIO;i,j}$ base their decision towards $S^{\oplus}$ or $S^{\ominus}$ on a K Nearest Neighbour (KNN) classifier [34, 218], previously trained on an initial set of audio profiles which notably do not include explicit person-specific information, but instead represent general settings such as indoor or outdoor scenarios [234].

Due to the limited length of the recordings it was decided to forego a meticulous modeling of trust in favour of the fusion of the agents' subjective opinions. The first part hence evaluates two variants V_1 and V_2 for models comprised of only the $L^i_{GEO;i,j}$. Each variant differs in its mapping from probabilities $p^{\oplus}$ and $p^{\ominus}$ to binary opinions $\omega^i_{GEO} = (b, d, u, a)$, where $p^{\oplus}$ and $p^{\ominus}$ denote the probabilities with which A_i and A_j do or do not interact. The first variant defines the mapping

$$V_1 : \left(p^{\oplus}, p^{\ominus}\right) \mapsto \left(b = p^{\oplus}, d = p^{\ominus}, u = 1 - p^{\oplus} - p^{\ominus}, a = \frac{1}{2}\right) \tag{169}$$

for which a was chosen as the default base rate [162] instead of any prior probabilities based on previous interactions for each pair of A_i and A_j. The second variant enforces a rigid *uncertainty boundary* through a constant value of $u = \frac{1}{4}$. This value was heuristically chosen to model remaining uncertainty based on the overall accuracies from the previous evaluation of various classifiers (see table 10 on page 70). It will also be justified *a posteriori* by the final evaluation results (see table 32). As a consequence of setting $u = \frac{1}{4}$, belief and disbelief have to be chosen according to equation (149 on page 171):

$$V_2 : \left(p^{\oplus}, p^{\ominus}\right) \mapsto \left(b = \frac{3}{4} \cdot \frac{p^{\oplus}}{p^{\oplus} + p^{\ominus}}, d = \frac{3}{4} \cdot \frac{p^{\ominus}}{p^{\oplus} + p^{\ominus}}, u = \frac{1}{4}, a = \frac{1}{2}\right) . \tag{170}$$

Now, let $\omega^{ij}_{GEO} = \omega^i_{GEO;i,j} \,\underline{\oplus}\, \omega^j_{GEO;j,i}$ and $\omega^i_{GEO} = \omega^i_{GEO;i,j}$, both layer II sensors which output their belief in $S^{\oplus}$ (as opposed to $S^{\ominus}$) as the result of applying some decision function to the outputs of either the fusion of several or just single sensors.

Table 32 shows the evaluation results for both variants and three choices of decision functions, according to which f_1 shows the best performance. Both f_2 and f_3 are more strict than f_1 since both require $p^{\oplus}$ to be significantly higher than $p^{\ominus}$ for a decision towards $S^{\oplus}$. For the present dataset, the difference is about 6% and therefore neglible. The choice of f_1 is further sustained by the fact that both f_2 and f_3 achieve higher precision in general, albeit at the expense of recall. The accuracy of the classifiers is roughly equal for each variant and choice of decision function. Comparison of the results however yields the most homogeneous performance for f_1 regardless of V_1 and V_2.

Among V_1 and V_2 the latter exhibits the overall better results. Accuracy and precision are both higher while recall is roughly the same for all configurations. It is interesting to have a look at the uncertainty component in case of V_1. Resulting from the underlying model of interaction geometry, the values for $p^{\oplus}$ and $p^{\ominus}$ can be fairly small for certain

Decision function	What	Variant	Accuracy	Precision	Recall
$f_1 = \begin{cases} S^{\oplus} & \text{if } b \geqslant d \\ S^{\ominus} & \text{else} \end{cases}$	$\boldsymbol{\omega}^{ij}_{GEO}$	V_1	0.731	0.676	0.764
		V_2	0.761	0.721	0.758
	ω^{i}_{GEO}	V_1, V_2	0.757	0.719	0.748
$f_2 = \begin{cases} S^{\oplus} & \text{if } b - d > \frac{d}{2} \\ S^{\ominus} & \text{else} \end{cases}$	$\boldsymbol{\omega}^{ij}_{GEO}$	V_1	0.739	0.724	0.669
		V_2	0.772	0.796	0.656
	ω^{i}_{GEO}	V_1, V_2	0.760	0.768	0.664
$f_3 = \begin{cases} S^{\oplus} & \text{if } b - d > \frac{b+d}{2} \\ S^{\ominus} & \text{else} \end{cases}$	$\boldsymbol{\omega}^{ij}_{GEO}$	V_1	0.723	0.802	0.503
		V_2	0.746	0.874	0.502
	ω^{i}_{GEO}	V_1, V_2	0.736	0.841	0.502

Table 32.: Classification performance based on opinions of logical GEO sensors under varying mappings and decision functions. Precision and recall were computed with respect to $S^{\oplus}$.

input parameters $(\delta\theta, \delta\varphi, dd)$. Such a configuration would leave relatively large values for the uncertainty u in V_1, and most likely have deteriorating influence on the results of sensor fusion like in equation (165). The proposed system therefore favours V_2 over V_1 as a mapping from probabilities to binary opinions.

Most importantly, the performance evaluation shows that in all cases the fusion of two sensors $\boldsymbol{\omega}_{GEO}$ yields better performance than ω_{GEO} alone. This result is further corroborated by the evaluation of the fusion of *distinct* kinds of level II sensors, for which

$$\boldsymbol{\omega}_{AUDIO} = \omega^{i}_{AUDIO;i,j} \underline{\oplus}\, \omega^{j}_{AUDIO;j,i} \tag{171}$$

and

$$\boldsymbol{\omega}_{GEO\underline{\oplus}AUDIO} = \left(\omega^{i}_{GEO;i,j} \underline{\oplus}\, \omega^{j}_{GEO;j,i}\right) \underline{\oplus} \left(\omega^{i}_{AUDIO;i,j} \underline{\oplus}\, \omega^{j}_{AUDIO;j,i}\right) . \tag{172}$$

The results are given in table 33. As expected, comparison of the performances of either one of $\boldsymbol{\omega}_{GEO}$ and $\boldsymbol{\omega}_{AUDIO}$ with that of $\boldsymbol{\omega}_{GEO\underline{\oplus}AUDIO}$ shows that SL fusion of the distinct subjective opinions of the agents yields a notable benefit for algorithmic inferral of social situations.

Decision function	What	Variant	Accuracy	Precision	Recall
$f_1 = \begin{cases} S^{\oplus} & \text{if } b \geqslant d \\ S^{\ominus} & \text{else} \end{cases}$	$\boldsymbol{\omega}_{AUDIO}$	V_2	0.737	0.709	0.695
	$\boldsymbol{\omega}_{GEO\underline{\oplus}AUDIO}$	V_2	0.785	0.756	0.767

Table 33.: Classification performance based on opinions of single and fusioned logical AUDIO and GEO sensors. Precision and recall have been computed with respect to $S^{\oplus}$. Table taken from [125].

4.6.1 *Evaluation of clustering with or without sensor fusion*

Following the discussion from section 4.5 the A_i build their personal opinions $\omega^i_{\{i,j\}}(t)$ and also request the $\omega^k_{\{k,l\}}$ from other nearby agents A_k, where $k, l \neq i$. Based on the collection of their own and the other opinions, a probabilistic view of the situational SN is won by clustering the graph $G(t) = (V, E, w)$ whose vertices represent the agents and whose weighted edges represent the probability of two agents sharing one social situation. More precisely, the weights $w(e_k, e_l)$ correspond to the expected probabilities of the $\omega^i_{\{k,l\}}(t)$ under a default base rate of $a = 0.5$ (refer to equation (156). Since G is based on mutual agreement, the lack of an edge between two vertices is equivalent to a zero-weighted edge. For comparison, all of Single Link, Complete Link and Average Link clustering [102] were evaluated. For each approach the optimal height of the corresponding dendrograms was determined according to Maximum Modularity [221, 220]. Greedy Maximization of Modularity [220] which is derived from Dijkstra's single-source shortest-path algorithm was evaluated as well. Maximum Modularity is defined as

$$\mathcal{Q} = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \delta(c_i - c_j) \right) , \tag{173}$$

for which $A_{ij} = w(e_i, e_j)$ denotes the weight of the edge between agents i and j, $k_i = \sum_j A_{ij}$, $m = \frac{1}{2} \sum_{ij} A_{ij}$, and c_i denotes the index of the community (here: social situation) to which i is assigned. δ is the delta function $\delta(x) = 1$ for $x = 0$, else 0. As such, $\mathcal{Q}$ corresponds to the number of edges within social situations minus the expectation in a random graph, given the assigned communities c_i and the A_{ij}.

The results of the clustering process were compared with the manual annotation of the social situations from section 2.2.3, and performance was measured in terms of the Rand index [260] and the *adjusted* Rand Index [147] which are defined as follows: Let $\mathcal{C}$ and $\mathcal{C}'$ distinct clusterings of the same $G(t)$, and let k and l the number of clusters under $\mathcal{C}$ and $\mathcal{C}'$. Furthermore, let N denote the total number of vertices. The Rand Index

$$\mathcal{R}\left(\mathcal{C}, \mathcal{C}'\right) = \frac{a + b}{\binom{N}{2}} \qquad \text{with} \qquad 0 \leqslant \mathcal{R}\left(\mathcal{C}, \mathcal{C}'\right) \leqslant 1 \tag{174}$$

measures the relation between a the number of vertices in the same cluster and b the number of vertices in different clusters under $\mathcal{C}$ and $\mathcal{C}'$. For large N the index will converge towards 1 due to the increasing number of clusters, in particular those consisting of only a single vertex. The Adjusted Rand Index therefore takes into account the expected value of the index under a generalized hypergeometric distribution [147, 338], i. e.

$$\mathcal{R}_{adj}\left(\mathcal{C}, \mathcal{C}'\right) = \frac{\mathit{Index} - \mathit{Index}_{Exp}}{\mathit{Index}_{Max} - \mathit{Index}_{Exp}} = \frac{\sum_k \sum_l \binom{m_{kl}}{2} - t_3}{\frac{1}{2}(t_1 + t_2) - t_3} \tag{175}$$

where

$$t_1 = \sum_k \binom{|\mathcal{C}_i|}{2} , \quad t_2 = \sum_l \binom{|\mathcal{C}'_j|}{2} , \quad \text{and } t_3 = \frac{2 t_1 t_2}{\binom{N}{2}} , \tag{176}$$

	$< \mathcal{R}\left(\mathcal{A}(t), \mathcal{C}(t)\right) >_t \pm\sigma_t$			$< \mathcal{R}_{adj}\left(\mathcal{A}(t), \mathcal{C}(t)\right) >_t \pm\sigma_t$		
Algorithm	*GEO*	*AUDIO*	$\mathit{GEO}\underline{\oplus}\mathit{AUDIO}$	*GEO*	*AUDIO*	$\mathit{GEO}\underline{\oplus}\mathit{AUDIO}$
AvL	0.77 ± 0.20	0.74 ± 0.31	0.78 ± 0.22	0.53 ± 0.37	0.57 ± 0.42	0.57 ± 0.39
SiL	0.75 ± 0.22	0.69 ± 0.34	0.78 ± 0.26	0.51 ± 0.37	0.54 ± 0.44	0.60 ± 0.39
CoL	0.76 ± 0.20	0.74 ± 0.31	0.78 ± 0.22	0.52 ± 0.38	0.58 ± 0.43	0.56 ± 0.39
GrM	0.76 ± 0.21	0.74 ± 0.29	0.77 ± 0.21	0.57 ± 0.40	0.56 ± 0.40	0.55 ± 0.39
Random	$< \mathcal{R}\left(\mathcal{A}(t), \mathcal{C}_{random}(t)\right) >_t = 0.524 \pm 0.233$			$< \mathcal{R}_{adj}\left(\mathcal{A}(t), \mathcal{C}_{random}(t)\right) >_t = 0.022 \pm 0.181$		

Table 34.: Rand and Adjusted Rand Indexes for combinations of single and fusioned sensors for Average Link (AvL), Single Link (SiL), Complete Link (CoL), and Greedy Maximization of Modularity (GrM).

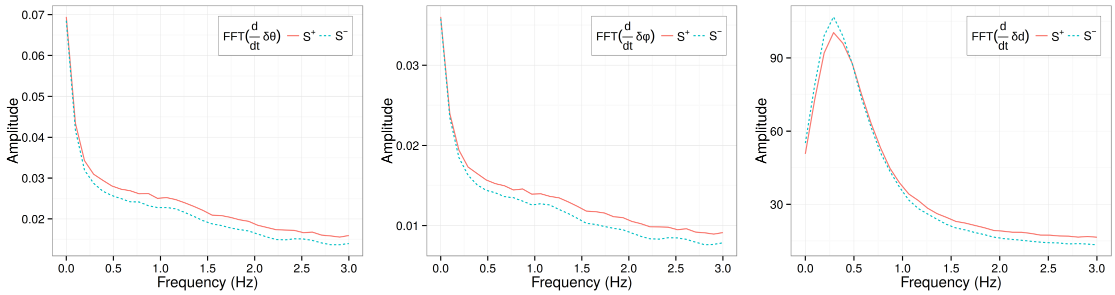
and the number of vertices in the intersection of $\mathcal{C}_k$ and $\mathcal{C}'_l$ is denoted by $m_{kl} = |\mathcal{C}_k \cap \mathcal{C}'_l|$. The final evaluation compares the clustering performances of $L^i_{GEO;SS}$, $L^i_{AUDIO;SS}$, and $L^i_{GEO \oplus AUDIO;SS}$, i. e. those top-level II logical sensors that output the current situational SN as seen by agent A_i based on either single or fusioned sources of information. The results are shown in table 34. It follows that the fusion of level Ib logical sensors yields significant better results after clustering than any of the sensor alone. For example, a Wilcoxon Rank Sum Test rejected both of the hypotheses $\mathsf{Median}\left(\mathcal{R}_{AvL;GEO \oplus AUDIO}\right) = \mathsf{Median}\left(\mathcal{R}_{AvL;GEO}\right)$ and $\mathsf{Median}\left(\mathcal{R}_{adj;SiL;GEO \oplus AUDIO}\right) = \mathsf{Median}\left(\mathcal{R}_{adj;SiL;AUDIO}\right)$ with significance level $\alpha = 0.05$. This is further sustained by a two-sided T-test rejecting $\mu_{GEO \oplus AUDIO} = \mu_{GEO}$ and $\mu_{GEO \oplus AUDIO} = \mu_{AUDIO}$ for the same confidence interval [125]. It should be noted that in some cases the results based on *AUDIO* alone were better than those of $\mathit{GEO} \oplus \mathit{AUDIO}$ (table 34). This was however explained after thorough analysis of the dataset in which it was found that during a relatively long social situation among all interactants of the experiment (table 2), a high number of frames yields artificats in terms of high variance of the interaction probabilities based on *AUDIO*. This effectively leads to a maximum in modularity for a single cluster, which just happens to coincide with the real SN at that very moment. This is the case for about 2% of all recorded frames. Eventually, leaving out the respective frames yields better results for *all* of Single Link, Average Link, Complete Link and Greedy Maximation of Modularity.

5 CO-ACTIVITY DETECTION

5.1 MODELING DYNAMIC SITUATIONS

The hitherto discussed approach for algorithmical models for social situation detection is based on the assumption that human behaviour is to a certain extent generalizable. A number of evaluations and discussions sustain this notion particularly with respect to social interaction geometry. It has also become clear that a multitude of known and unknown variables may affect the model, such as e. g. gender or group size. The proposed model has still proven to be rather robust and universally applicable throughout corresponding experiments with controlled and uncontrolled variations of the aforementioned parameters. On the other hand, there are undoubtedly situations where a static model of interaction geometry is prone to fail. Consider for example a ride on the subway. If the train were packed with lots of people, how could the proposed model be used to achieve reliable results on who is interacting with whom? Also, what would happen if instead there were less people on the train, but had to sit close together and possibly face each other? What about visiting a theater, attending a rock concert, or dancing at the Vienna Opera Ball? A model based solely on point estimates of interaction geometry is likely to fail due to both *static* and *dynamic* components, which neither *are* nor possibly *can* be considered by the model in its current form. The knowledge of the fact that individuals work together in order to uphold established spatio-orientational arrangements [28, 114, 166] may compensate for a limited number of dynamic components with bounded magnitude, but the same cannot apply to all dynamic components in general.

One possibility to overcome a number of those problems could be to construct a model either based on more than just interaction geometry, as was shown in chapter 4 where the SL fusion of logical sensors of interaction geometry and of low-level audio features led to significantly improved performance, or considering more than just independent observations, each of which merely represents a single point in time. This kind of approach would require to embed samples in their relevant context, be it social context, such as when feeding the model *a priori* information about social relations, or be it timely context, such as when dealing with sequences of observations. In other words, history-based estimates of social situations may naturally lead to improvements over point-based estimates. The beginning of this chapter therefore discusses a number of possible alternatives to static analysis as context for the introduction of the newly proposed model for co-activity detection.



(a) (b) (c)

Figure 49.: Amplitude spectra of sequential changes in $\delta\theta$, $\delta\varphi$ and δd. The spectra were computed based on a sampling frequency of $Fs = 6Hz$ and using a sliding $10s$ Hamming window with $5s$ overlap.

5.1.1 *Analyzing the frequency domain*

One may be inclined to analyze just how much the values of $\delta\theta$, $\delta\varphi$ and δd change over time. In the presence of social interactions one would consequently expect rather small adaptions, i. e. notably less entropy. It seems however difficult to say whether the absence of interaction will be guaranteed to yield a different picture. To see this, the amplitude spectra of the derivatives of $\delta\theta$, $\delta\varphi$ and δd with respect to time can be investigated for both $S^{\oplus}$ and $S^{\ominus}$. The derivatives are computed for sliding windows of equal length over each variable. Before transforming the signal from the time to the frequency domain the samples are weighted using a Hamming window in order to counteract effects like leakage or additional high frequency components [359, 304] as the signal is actually not periodic. Each window (*frame*) f^k is then transformed into the frequency domain using the Fast Fourier Transform (FFT), yielding a vector F^k whose elements correspond to the respective frequency components. From these vectors of complex numbers the signal's amplitude r and phase ϕ are computed as

$$r(F_i^k) = \sqrt{\mathrm{Re}(F_i^k)^2 + \mathrm{Im}(F_i^k)^2} \quad \text{and} \quad \phi(F_i^k) = \mathrm{atan2}\big(\mathrm{Im}(F_i^k),\, \mathrm{Re}(F_i^k)\big), \tag{177}$$

where F_i^k denotes the i-th component of the vector F^k in the frequency domain. Note that the frequency resolution depends on the chosen window size [359, 304]. The original signal was sampled at a frequency of $Fs = 6$ Hz (refer to section 2.2). Selecting a window size of e. g. 10 s yields $\lfloor 10s \cdot Fs/2 \rfloor + 1 = 31$ distinct frequency bins, ranging from 0 Hz to 3 Hz (the factor 2 being explained through symmetry of the F_i^k for negative frequencies). Averaging over the resulting spectra of all pairs of persons results in the final mean amplitude spectra, corresponding to either $S^{\oplus}$ or $S^{\ominus}$. These are depicted in figure 49.

As expected, the spectra are quite distinct for each variable and class. It can be assumed that series of $\delta\theta$, $\delta\varphi$ and δd are well separable according to their actual class, possibly by using linear models such as logistic regression or SVMs. It is interesting though that in spite of their actual differences the spectra for $S^{\oplus}$ and $S^{\ominus}$ are still quite similar for each variable. It is supposed that this is again, at least partially, a consequence of the spatial constraints during the original experiment (see chapter 2). As opposed to the model for *static* interaction geometry, it constrains the possible *changes* of the variables' values in regard of dynamic analysis. It is nevertheless suspected that in particular the higher frequency components of the spectra of δd and $\delta\varphi$ will exhibit more significant differences once more data are gathered.

5.1.2 *Hidden Markov Models*

The insight that for a given pair of persons the presence or absence of social interaction at time t yields a high probability of remaining in the same social state at $t+1$ eventually leads to the notion of Markov chains which may be used to model changes in the value of a random variable X over time. The order M of a Markov chain defines the conditional probability $p(x_t|x_{t-1},\dots,x_{t-M})$. In other words, the probability of observing x at time t depends only on its M previous instances. It follows that the joint distribution of observing a sequence of N values is given by

$$p(x_1,\dots,x_N) = p(x_1)\prod_{t=2}^{N} p(x_t|x_{t-1},\dots,x_{t-M}) \,. \tag{178}$$

A M-th order Markov chain, corresponding to a discrete random variable with K states, is defined by $K^{M-1}(K-1)$ independent parameters [34, 218]. As computations might otherwise become intractable, the dependency assumption is usually relaxed to the first order [34]. A related form of Dynamic Bayesian Networks (DBNs), namely HMMs, follow this approach by introducing a set of latent variables. Each observation x_t is accompanied by a corresponding latent variable z_t [34, 218]. The z_t are defined as discrete random variables, and instead of the x_t now the z_t form a first-order Markov chain. Conditioning the x_t on their corresponding z_t gives rise to the joint distribution

$$p(x_1,\dots,x_N,z_1,\dots,z_N) = p(z_1)\left[\prod_{n=2}^{N} p(z_n|z_{n-1})\right]\prod_{n=1}^{N} p(x_n|z_n) \tag{179}$$

for observing the corresponding sequences of values of X and Z. It can be shown that x_t in fact depends on all its previous observations [34]. The z_t are known as *state* variables. Observations are therefore probabilistic functions of state [256, 255]. Also note that observations can correspond to both discrete or continuous random variables. As such, HMMs can be seen as a generalization of mixture models whose components are not i.i.d. but instead follow a Markov process [34]. Another important property of HMMs is that they are to some extent considered to be invariant to compression or streching of the time axis [34]. HMMs are used for numerous applications e. g. in speech recognition, handwriting recognition, activity recognition, or DNA analysis [34, 256, 103, 241]. As generative models, they are commonly used for the prediction, filtering, smoothing, and classification of sequential data. According to Rabiner [255], the three most relevant problems that HMMs solve are:

- Determining how well a particular model fits a given sequence of observations.
- Computing the most likely sequence of (hidden) states for a given sequence of observations.
- Maximizing the probability of observing a given number of sequences of observations.

In regard of sequences of observations of $\delta\theta$, $\delta\varphi$ and dd, some or even all of the above items also apply to the present problem domain. This may however depend on the particular model design and/or choice of parameters. In [122], Groh and Lehmann show the results of evaluating a model with only two hidden states corresponding to $S^{\oplus}$ and $S^{\ominus}$. Either one of GMMs or quantized training data were used for the distributions of the variables as observable from each state. Evaluation of the model on the R2B dataset (see section 2.3.5.6) resulted in a classification accuracy of about 74%, which is about the same as for the *static* model of interaction geometry. This is however not surprising as the states directly correspond to the classes and were also observable from the training data. Hence the model should be considered as a first-order Markov chain rather than a HMM. It also means that the model merely added state transition probabilities in comparison to the previous model. Interestingly enough, it furthermore turned out that any choice of the initial state probabilities according to $\pi = (i, 1 - i)$ for $i \in [0, 1]$ had no relevant impact on the results which speaks for a rather smooth surface of the optimized function due to the distribution of the data and stable convergence characteristics of the model.

It is trivial to see that the choice of the number of states is crucial and that this is also interdependent with the choice of probability distributions for the observed variables. This is a typical design problem for HMMs, for which Rabiner suggests an iterative process [256, 255]. That process means making an initial choice of model parameters, followed by computing the most likely sequence of hidden states for a given sequence of observations, and subsequent analysis of pairs of corresponding states and observations. This may lead to an understanding of why which observation was assigned to which state, and possibly give an idea of how that should affect e. g. the number of states or choice of probability distributions. Despite of or in addition to this tedious process several options come to mind: The initial number of states could be chosen according to heuristically determined sectors in $\delta\theta$, $\delta\varphi$ or δd. Visual inspection of the experimental datasets has shown that the data are distributed among several clusters (see section 2.2.5). Doing so may also provide a way to incorporate additional parameters such as group size, gender or age. Recall that the variables' distributions were quite distinct for varying values of those parameters.

An alternative could be to make random (or controlled) choices for the number of states and using model selection to figure out which model suits best. This approach is considered disadvantageuous not only due to the high computational efforts but particularly so because it will likely result in an overfitted model. Nevertheless, it may still be advantageous due to the fact that a random choice for the number of states would imply neither accidental nor explicit insertion of heuristics into the model. Making a random choice and then using EM-based learning, such as the Baum-Welch algorithm [34, 255], might then even lead to discovering previously unseen patterns in the distribution of the data among the states.

It should be noted that although it seems reasonable to use GMMs or SW-GMMs for the probability distributions of the variables as observed from each state, depending on the number of states and the consequent actual distribution of the observed values per state it may be worth considering other distributions. Recall that one of the reasons why mixture models were favoured so far was due to the clear presence of clusters along with the observable variance in the data. This need not be the case once EM leads to a different distribution of the data on a per-state basis. To the contrary, overly complex models might lead to overfitting, resulting in disproportionally high likelihoods of some observations and thus the corresponding states. In fact, the results in [122] indicate that quantization of the data may be sufficient for HMMs temporal analysis of dynamic social interaction data.
In regard of classification, one must also consider whether classification should be based on a single or two separate HMMs. If a single model were to be used, the states had to be partitioned into two sets, each of which correspond to either one of $S^{\oplus}$ or $S^{\ominus}$. The number of states per class need not necessarily be the same. After computing the most likely sequence of states for a given series of observations, that state sequence ought to be smoothened. Majority voting would finally yield the classification result. If two separate models were to be used, however, states would correspond to neither class. Instead, each of the models were trained per class, and classification would be done by deciding for the class of the model with the higher likelihood of observing the given sequence. Note that using several models is common practice in e. g. speech or handwriting recognition [34, 256]. As a matter of fact, further research of this particular topic is beyond the scope of this thesis. At the end of the next section, additional reasons will be given for why a different approach was chosen.

5.1.3 *Eigenzone decomposition*

In 2009, Eagle and Pentland published a seminal work on representing routine behaviour in terms of a set of characteristic vectors, together describing the so-called eigenbehaviour of entities such as single persons or groups [83]. Their work is based on previous research by Turk and Pentland who used a similar approach for application in face recognition for which the characteristic vectors were formerly known as eigenfaces [328, 329].
In both cases the basic idea is the representation of complex data as a weighted sum of its principal components, determined as the eigenvectors of the covariance matrix of the original data. Following the insight that high-dimensional data are typically not just randomly distributed but instead can be described by a lower dimensional space [329], it was shown that the major part of the data's variance could in fact be explained through a small number of eigenvectors. Hence a dataset is effectively reduced by projecting it into its corresponding eigenspace. Identification of behaviour, or recognition of faces, can subsequently be achieved e. g. by means of clustering or KNN search. In case of face recognition, for example, a set of 114 65536-dimensional images (256 by 256 grey pixels) could be exhaustively explained through merely 40 eigenfaces [328]. As the eigenfaces form a common basis for all images, information in each image could thus be represented by $\sim 2^6$ instead of 2^{16}

coordinates. For the eigenbehaviour problem, on the other hand, Eagle and Pentland used their well-known Reality Mining dataset [82] for the analysis of social routine behaviour. This dataset is comprised of 9 months worth of recording rich data from the mobile phones of 100 subjects, such as "location, proximate phones, and communication" [82] (see chapter 1). Analysis of the principal components puts emphasis on the variance in a person's daily behaviour while neglecting average behaviour. Given a set of M-dimensional vectors $\Gamma_1, \dots, \Gamma_N$, each corresponding to one day's recordings of M variables, Eagle and Pentland combine subsequent vectors to represent longer periods of time. This way, a matrix of $N \times M$ daily samples can be transformed into arbitrary representations of $\frac{N}{d} \times Md$ matrices, where d denotes the chosen number of days. Eagle and Pentland found that about six eigenvectors would be sufficient to represent a person's eigenbehaviour, of which the most important aspect turned out to be location [83]. Interestingly enough, they note that those six eigenvectors would describe "individuals within the business school community with 90% reconstruction accuracy, but the senior lab students with 96% accuracy" [83]. This implies that aside from information loss through dimensionality reduction by means of the selected number of eigenvectors there is probable cause that the recorded variables were themselves not capable of exhaustively describing social behaviour, or capturing all of the necessary social context, which is not unexpected.

The eigenbehaviour principle is by all means transferable to the present problem domain. Subsequent samples of $\delta\theta$, $\delta\varphi$ and δd, representing a window of N seconds, can be concatenated to a single $3N$-dimensional sample. All of those samples from both $S^{\oplus}$ and $S^{\ominus}$ can then be collected in a single matrix D, and the eigenvectors of the covariance matrix of D be determined via numerically stable SVD. The eigenvectors would consequently describe temporal *eigenzones*, and span a subspace into which the zero-mean data can then be projected. KNN or similar algorithms could then be used for the discrimination of newly recorded samples between $S^{\oplus}$ and $S^{\ominus}$. All the same, the implementation and evaluation of this approach is left as an open question. While it is assumed that eigenzone decomposition of the data will yield reasonable classification performance, it seems that a corresponding model would be rather restricted, e. g. in terms of being dependent on heuristic choices of the window length N in accordance with a respective application domain. As an example, the model might be able to explain social interaction at the Vienna Opera Ball, since the movement of a dancing couple relative to each other differs very much from their movements relative to other couples, yet the very same model will likely fail to "understand" a group of people playing soccer. Likewise problems apply to a number of more or less related approaches like Goldberger's Neighbourhood Component Analysis (NCA) [116], representing the data by means of Non-Negative Matrix Factorization (NMF) [186] based on specifically selected or designed components, or finding other common properties with respect to supposedly lower-dimensional manifolds, such as by Bishop and Svensen's Generative Topographic Mapping (GTM) [33, 35].

5.2 THE PROPOSED MODEL

All of the previously discussed techniques for modeling dynamic social situations approach the problem from a different angle, albeit mostly in terms of mere different representations of the data so that some might work temporarily better whilst others will not. None, however, explicitly add something significantly new to the cause. Generally speaking, all of them are susceptible to careful selection of the model parameters with respect to the concrete application domain. Without doubt, this will always imply a non-negligible constraint through certain heuristics. In spite of the assumption that social behaviour can be generalized, even if only to a certain extent, the resulting models are prone to overfitting. Note that in this context the term “overfitting” is not limited to the sense of overfitting a particular dataset. Instead, it extends to the notion of being restricted to *that particular portion* of the original problem domain that was understood when the respective model was designed. It seems unlikely though that social behaviour can be grasped to an extent that makes it possible to *fully* understand any particular domain. This is e. g. supported by Lane et al. according to whom “mobile phones are often used on the go and in ways that are difficult to anticipate in advance. This complicates the use of statistical models that may fail to generalize under unexpected environments” [181], and further that “anticipating the different scenarios the phone might encounter is almost impossible” [181].

For the purpose of detecting whether two persons interact it therefore makes sense to exploit this particular insight and consequently reduce the amount of explicitly and implicitly introduced heuristics to a potential minimum. The interpretation of physical signals and/or sensations obviously makes sense for humans and machines alike. For a machine learning model, abstracting from raw data makes further sense for two reaons: The model might otherwise be intractable, both analytically and/or computationally, and it might be impossible to interpret or understand the model itself. Whereas humans in principle have continuous access to both their original raw sensations as well as their (logical) interpretations of the former, for machine learning models it is almost inevitable that interpretation of the raw data in terms of features goes along with a reduction of information. Since trained models are naturally bound to the information they have “seen”, an alternative idea for the detection of dynamic social interaction can thus be outlined as the pairwise analysis of concurrent datastreams from mobile sensors belonging to the subjects in question whilst refraining from interpreting those data as much as possible. In other words, the detection of conjoint patterns in concurrent datastreams is equivalent to detecting whether two or more subjects perform the same type of activity simultaneously. Since knowledge of the precise type of activity (e. g. running, playing soccer, cooking, etc.) is irrelevant for the detection of mutual interaction, a corresponding model is deemed much more generalizable than other models from the field. In the following, this approach will lead to the concept of *co-activity detection* as a new contribution to the broader fields AR and SSP.

5.2.1 *Activity Recognition*

The *detection* of activities as such is part of AR. Since its advent in the late 1990s, it has gained much interest along with the substantial advances in pervasive computing and mobile sensing. Not surprisingly, AR has many applications ranging from academic research to personal and environmental monitoring, rehabilitation, health and elderly care, emergency help, performance sports, social networks, business and transportation [16, 327, 181, 230]. Another interesting aspect is the processing of enormous datasources, e. g. in online media, where AR in videos is required for automatic “content-based video annotation and retrieval, highlight extraction and video summarization” [327]. According to Avci et al. [16], “the goal of activity recognition is to recognize the actions and goals of an agent or a group of agents from the observations of the agents’ actions”. Turaga et al. further distinguish between *primitive actions*, also known as *atomic actions*, which may occur in a single instant or even take up to a few seconds, and which are subsumed by the more complex *activities*, the latter of which represent coordinated actions [327]. Note that both entities can be interpreted as words and sentences of a language. Indeed a number of algorithmic approaches for AR are based on the use of *grammars* [327]. In terms of time series of observed variables, Lara and Labrador [182] define the Human Activity Recognition Problem (HARP) as follows:

> “Given a set $S = \{S_0, \ldots, S_{k-1}\}$ of k time series, each one from a particular measured attribute, and all defined within time interval $I = [t_\alpha, t_\omega]$, the goal is to find a temporal partition $< I_0, \ldots, I_{r-1} >$ of I, based on the data in S, and a set of labels representing the activity performed during each interval I_j (e. g. sitting, walking, etc.). This implies that the time intervals I_j are consecutive, non-empty, non-overlapping, and such that $I = \bigcup_{j=0}^{r-1} I_j$.”

It is rather obvious that AR, being mostly based on supervised learning of statistical models, faces the same problems that were already mentioned several times throughout this thesis. For example, Avci et al. report that “differences between cultures and individuals result in variations in the way that people performs tasks” [16]. In addition to that, yet another problem is given by the hierarchial organization of activities. [167] for example phrases this as follows: “Individual acts, moment to moment behavioural events, are not just concatenated together, one after the other, but are always under the guidance of a larger plan of some sort.” This may go as far as people performing multiple tasks at once. Together, the aforementioned issues cause a multitude of problems for both, the segmentation of a stream of activities, as well as their precise recognition. [181] eventually conclude that “existing statistical models are unable to cope with everyday occurrences such as a person using a new type of exercise machine, and struggle when two activities overlap each other or different individuals carry out the same activity differently”. Steele et al. therefore propose to regard manual annotations only as hints instead of absolute truth, [311] in [181]. Likewise, *active learning* utilizes initial labels from the training data as soft guesses [144]. In regard of semi-supervised and unsupervised approaches, Poppe [248] notes that “when no labels are available, an unsupervised approach needs to be pursued but there is no guarantee that the discovered classes are semantically meaningful”.

5.2.2 *Co-Activity Detection*

Arguably, the proposed approach of co-activity detection is more generalizable in the sense that it aims at *detecting* social co-activities, but *not necessarily recognizing* the exact type of activity that was performed. For this, the terms activity, co-activity, deferred co-activity, and social co-activity are defined as follows [19]:

- An *activity* is described by a four-tuple (S, T, X, K), where S denotes a singleton whose only element refers to the person who is performing the activity, $T \in \mathbb{R}$ references the time at which the activity was performed, $X \in \mathbb{R}$ references the location at which the activity was performed, and K with $|K| \geqslant 0$ is a set of tags which sufficiently describe the action’s semantics. Note that this definition is close to the definition of social situations from chapter 1, except for the singleton P. Accordingly, T and X are projections from a spatio-temporal reference $\tilde{X} \in \mathbb{R}^4$.
- A *co-activity* is described by a four-tuple (P, T, X, K). In this case, P references a set of persons, subject to $|P| \geqslant 2$, who perform the exact same activity, which is described by K, at time T and location X. This does not imply that all persons in P need to be mutually aware of each other.
- A *deferred co-activity* relaxes the former definition of co-activity as follows:
  - A *spatially deferred co-activity* corresponds to a co-activity performed by a set P of persons at time T, but not necessarily at the same location. It is thus described by a three-tuple (P, T, K), subject to $|P| \geqslant 2$.
  - A *temporally deferred co-activity* corresponds to a co-activity performed by a set P of persons at location X, but not necessarily at the same time. It is thus described by a three-tuple (P, X, K), subject to $|P| \geqslant 2$.
- A *social co-activity* conforms to a co-activity plus the constraint that all persons in P are mutually aware of each other as well as the fact that they are performing the same activity. This awareness need not be conscious.

Note that including the descriptive set K of tags does not contradict the postulate that the proposed approach should primarily detect activities as such, but not necessarily recognize the exact types of those activities. For the detection of co-activities, it is sufficient to use abstract tags, provided that they allow for the distinction of different activities as such.

Deducing evidence for short-term as well as long-term social relationships is arguably not bound to labeled activities. Although, generally speaking, precisely knowing the activities’ types would allow for a deeper understanding of the subjects’ relationships, their interests etc., the mere knowledge that, when (how often, ...), and where activities were performed, together already yields substantial information as well. If, for example, two persons were to perform the same kind of activity on a regular basis, they might be training partners at sports or colleagues at work. If, on the other hand, different types of activities were to be performed on a short-term regular basis, that may hint towards very close friends or a spouse. It follows that in order to distinguish between the given examples, the interval and/or duration at which activities were performed has to be taken into account. For partners at sports, the duration would probably be less while intervals between instances would be longer, whereas for colleagues the contrary might hold. Any assumptions such as these will of course require future research. There is no doubt, however, that a grey zone will always remain, e. g. when training partners are simply colleagues at the same time, such as it might be the case for professional athletes, or because colleagues at some other business might go to the same fitness center after work.

In the following, emphasis is placed on the detection of co-activities, whereas deferred co-activities are rather considered as a by-product. To give at least a few examples for why those might be useful as well, evaluations of the latter might be helpful when there is a particular interest in a group of people who tend to perform the same activity at either the same time or the same location. Such knowledge could e. g. be employed for applications in surveillance, predicting the spread of diseases, or when a single party such as a manufacturing company wants to address a group of people with the same interests.

5.3 A FRAMEWORK FOR CO-ACTIVITY DETECTION

As part of the proposed framework for co-activity detection, a mobile application was developed in [19] to support continuous monitoring and recording of sequential datastreams from numerous mobile phone sensors. The application has been designed such that it is capable of recording all available types of sensors, but is also easily extendable to future sensor types. Unless explicitly chosen otherwise, sensory signals are recorded at the highest possible sampling rate and resolution, depending on what is supported by the mobile phone's operating system. The data are then stored on the mobile phone's flash drive in a compressed format. The software was developed for Apple iOS 6 because of its possibility to generate and deploy native code that would allow for frictionless recording of the relatively high bandwidth of data from the sensors. In comparison to other mobile platforms, iOS was deemed to support the least diverse hardware platforms which would likely allow for gathering unanimous experimental data from different subjects. Aside from access to the raw signals of most physical sensors, the operating system also provides a number of logical sensors as a result of sensor fusion. The output of these logical sensors is still rather low-level, which is why recording these signals is considered to be in agreement with the postulate of avoiding inclusion of explicit heuristics. As an example, operating system components like CoreLocation and CoreMotion support the fusion of several sensors in order to provide more precise estimates of device location and orientation. For this, CoreLocation may e. g. combine GPS, WiFi and ranges from mobile cell towers, whereas CoreMotion may combine three-dimensional input from the device's accelerometer, gyroscope and magnetometer. Orientation is expressed with respect to a particular reference frame. For this application, the reference frame was chosen such that the x-axis points towards true north while the z-axis points into the direction inverse to earth's gravitational force. In order to estimate true north as opposed to magnetic north, magnetic variance is taken into account depending on the device's current location, provided that information on the latter is available. Some operating system events, such as location updates, can be processed, or more specifically cached for further processing, when the application is in the background. This is unfortunately not true for all sensory input. Audio input streams, for example, although in principle a shared resource, can be cut off by the operating system and given to other foreground applications. Running the application in the foreground is therefore mandatory during experiments. With the recent advances in pervasive computing and taking into account development in e. g. mobile health monitoring systems, such as Apple's iWatch, it is however suspected that this constraint will vanish in the near future. Note that at the time of development iOS would not allow direct access to certain sensors, such as e. g. the proximity sensor which is normally used to turn off the display backlight when users are holding the phone against their cheek. Although not considered as a significant loss in comparison to the other available sensors, a proximity sensor, which basically works by evaluating the power of ambient light, could yield viable information about the phone's environment, and aid in the determination of the phone's on-body location. Further note that the operating system also does not permit arbitrary scans for SSIDs of nearby WiFi access points or unpaired Bluetooth devices in the vicinity. The following list gives an overview of the recorded data. Further details can be found in [19]:

- **Location**
  Estimates of the device's location are recorded in terms of latitude, longitude, and altitude. Course (°/s) and speed (m/s) may be recorded as well.

- **Proximity**
  Based on "found peer" and "lost peer" events from the operating system, other devices in the vicinity are detected via Bluetooth and/or WiFi connections. Note that operating system restricts these events to events from other iOS devices which run the same application concurrently. The Unique Device Identifiers (UDIDs) are then recorded for all such devices.

- **Compass and orientation**
  In addition to raw sensory input from the three-dimensional magnetometer, accelerometer, and gyroscope, enhanced (fusioned) readings are available as gravity (g), device acceleration (g), attitude (quaternion), and rotation rate (rad/s). Internally, band-pass filtering techniques are used to separate gravity and device acceleration components from the measured acceleration. Vice versa, the total acceleration equals the sum of gravity and device acceleration.

- **Audio**
  Monophonic audio is recorded at 8 KHz and compressed using the IMA 4:1 ADPCM codec. In spite of an undeniable loss in quality in comparison to 44 KHz and lossless encoding, these settings were chosen to reduce filesize and bandwidth when writing

to the device's internal flashdrive, thereby also avoiding possible IO related interrupts during long-time recordings.

- **Other**
  Device information and the current battery level are recorded, allowing for identification of a device's datastream as well as supplementary analysis of the energy consumption depending on the active set of sensors.

5.4 DATASET

In advance of the evaluation of the proposed system for co-activity detection, a dataset was collected, for which the sensor logging application was deployed on several iPhone 4 devices. Participants were asked to carry the phone in their right-hand front pocket of their trousers. No instructions were given regarding the phone's orientation. As placing the phone in the pocket would clearly deteriorate the quality of audio recordings, the participants each wore a headset (default iPhone headset), consisting of small headphones and a microphone built into the wire that connects the phone and headset. Initial synchronization of the devices was performed by means of using the phones' accelerometers to detect a shock. For this, participants would either place their phones on a common flat surface and one of them would then thump that surface, or participants would bump their phones together. This proved to be an efficient mode of synchronization, considering the nature and lengths of the recordings as opposed to the much more critical synchronization aspects during ultrasonic distance measurements which were discussed in section 3.3. In order to avoid explicit synchronization and thus involvement of the user, future approaches may want to investigate alternatives such as dynamic time warping [279], a technique well-established e. g. in speech recognition [257] which provides a "distance measure between two sequences, possibly with different lengths" [248], and thus the means to synchronize distinct sequences of observations.

Aiming at recording the preferably most natural behaviour of the subjects led to a selection of participants who were mostly not familiar with topics related to this work. To provide further guidance, *scriptlets* [19] were used to instruct the participants during the recording sessions. Individually or in combination, scriptlets outline an experimental scenario for the participants. It should be noted that in comparison to recording trials which were performed prior to the actual experiments, it turned out that the use of such scriptlets at least sometimes affected the participants' behaviour, of which some showed a tendency of acting less relaxed, most notably in phases of chatting. On the plus side of using scriptlets, however, they are less intrusive in regard of determining the ground-truth of the data. First, participants need not be actively involved in interactions with their mobile agents. Second, subsequent annotation by expert labelers will yield more congruency. This also means that problems arising from either the participants' subjective views on the activities, as well as possible hierarchical nesting of activities, can be avoided to a large degree. Table

35 provides a collection of all atomic scriptlets to be combined in arbitrary configurations, for example:

"GreetingStanding → WalkingTogetherIndoors → SittingDownTogether".

The final dataset consists of 34 clean recording sessions, i. e. sessions with proper time synchronisation and consistent sequential streams of sensor readings. In total, 6.7 hours were recorded over the course of a few weeks. 11 persons participated in the sessions, 4 of which were female. An average number of 3.4 sessions were recorded per subject (minimum 3, maximum 5), and durations varied from 4.3 min to 26.6 min per session, with a median and mean lengths of 10.9 min and 11.6 min, and a standard deviation of 5.3 min. Note that each session is comprised of several phases with varying co-activities and non-co-activities. Co-activities may be followed by other co-activities as well as non-co-activities, and vice versa. Annotation of the recorded sessions shows a clear bias towards co-activities with lengths of up to 5 minutes. Out of 6.7 hours total, 5.5 hours correspond to established co-activities and 1.2 hours to non-co-activities. Although not strictly necessary, co-activities were additionally labeled according to the scriptlets of the respective session.

5.4.1 *Postprocessing*

As mentioned before timely synchronization was performed by either placing the participants' phones on a flat surface and thumping on that surface, or by bumping the phones together. Doing so led to consistent peaks in the amplitudes of the devices' measured accelerations. The remainder of each recording therefore corresponds to the actual session, from which brief transitions after the peak and before the end of the recording were cut off consistently according to the respective session. Using the scriptlets in conjunction with the recorded audio streams allowed for manual annotation of the data corresponding to the prevalent activity, but most importantly according to whether co-activity was present ($C^{\oplus}$) or absent ($C^{\ominus}$). Annotations were always performed for pair-wise recordings. For this, a custom domain-specific language was used which allowed to define a default class (in this case $C^{\ominus}$), as well as to specify only those intervals which would differ from the default and how. This approach significantly simplified the efforts necessary for annotating the recorded sessions [19]. Figure 63 in appendix D illustrates the result of annotating a single session with two participants.

Next feature vectors were computed for the later classification of the data with respect to $C^{\oplus}$ and $C^{\ominus}$. Using a sliding window over each pair of concurrent datastreams in a session, numerous features were calculated for each window. The windows were centered around multiples of a chosen frame rate F_r. Varying values for the frame rate as well as window sizes were considered during the evaluation (see section 5.5). In addition to features describing the similarities or differences of pairwise streams, some features were also calculated for individuals streams. It will be shown that those features still occurred in semi-pairwise configuration in the resulting models (also discussed in section 5.5).

Recall that the proposed model should introduce as little explicit domain-knowledge as

Scriptlet	**Description**
GreetingStanding	Two persons meet and then greet each other while both of them are standing. The greeting can be just vocal, by shaking hands, or by hugging.
GreetingSitting	Two persons meet and then greet each other while one of them is sitting down while the other person is standing.
WalkingTogetherOutdoors	Two persons meet and then greet each other while one of them is sitting down while the other person is standing.
¬ WalkingTogetherOutdoors	Two persons walk around outdoors without interacting (no co-activity). While there is no co- activity the two persons are still physically close, for example one is walking behind the other.
¬ WalkingOutdoors	Two persons walk around outdoors without interacting. Their walking paths are not related or similar.
WalkingTogetherIndoors	Two persons walk next to each other inside a building and chat casually.
¬ WalkingTogetherIndoors	Two persons walk around inside a building with- out interacting (no co-activity). While there is no co- activity the two persons are still physically close, for example one is walking behind the other.
JoggingTogether	Two persons go jogging together and chat.
¬ JoggingTogether	Two persons go jogging and there is no interaction between them. Their paths are similar and they are physically close, for example one is jogging behind the other.
SittingDownTogether	Two persons sit down together and talk to each other.
¬ SittingDownTogether	Two persons sit down. While they are sitting next to each other there is no interaction between them (no co-activity).
ThrowingAndCatching	Two persons take turns throwing and catching a small object, for example a rubber ball.
DrivingTogether	Two persons drive together in the same car. One of them drives the car. They chat casually during the car ride.

Table 35.: Scriptlets used in the description of scenarios during the experimental sessions. Table taken from [19].

possible. In particular, features should not be based on concrete social or behavioural cues. An understanding of the latter might eventually be implicit as a result of learning the model. Therefore features such as turn-taking patterns or step frequencies were omitted. For each pair of devices, let the signal streams of each physical or logical sensor ψ be given as a function

$$f : \Psi, \mathbb{R}, \mathbb{R} \to \mathbb{R}^{2\times l'} : \psi, t, l \mapsto \begin{bmatrix} x \\ y \end{bmatrix} \tag{180}$$

of time t (in seconds) and window length l (in seconds), for which $l' = F^{\psi}\_s \cdot l$ corresponds to the number of samples depending on the sensor's sampling rate F_s^{ψ}. The following location-, motion- and audio-based features are computed as functions of the vectors $\mathbf{x}$ and $\mathbf{y}$, for which more details can be found in [19].

5.4.1.1 *Location-based features*

Distance between the two devices is computed from their measured (latitude, longitude, altitude) triplets. Although Euclidean distance can generally be considered a good approximation of the real distance between points A and B given in spherical coordinates, provided that A and B are sufficiently close together, that approximation deteriorates significantly with increasing magnitude of latitude. This feature is therefore computed as great-circle distance.

As location estimates are susceptible to numerous sources of noise and hence their quality may vary significantly [188], *location accuracy* is used as a feature for each device so as to provide the classifier with means of weighing other location-based features accordingly.

Recall that iOS does not allow direct scans for SSIDs of wireless access points or MAC addresses of arbitrary Bluetooth beacons in the vicinity. The sensor's ability to sense the presence of other devices is instead limited to those devices that run the same sensing application concurrently. Since the latter condition holds for all participants of the actual experiment, this *device proximity* feature is still useful. It is expressed as a boolean value indicating whether the two devices could "see" each other.

The last two location-based features are the *course delta*, describing the difference between the absolute courses of each device in degrees, and *speed delta*, describing their difference in speed in meters per second. The features are respectively based on the course and speed logical sensors of the phone, dependent upon the devices' location trails.

5.4.1.2 *Motion-based features*

Joint and separate features are computed from the devices' accelerations, rotation rates, and orientations. For each device, acceleration magnitude, three-axis base frequencies, gravity axis, and a number of simple statistical measures are computed.

Acceleration magnitude equals the magnitude of the medians of all three-axis acceleration measurements within the current window, i. e. $\left((m_x, m_y, m_z)(m_x, m_y, m_z)^T\right)^{\frac{1}{2}}$ where

m_x, m_y, m_z denote the medians of the values of the time series for the respective axes. The median was chosen to diminish the effect of outliers.

For each of the three measured axes, the corresponding *means*, *standard deviations*, *minima* and *maxima* are computed. In accordance with [280], the (minimum - maximum) and (maximum - minimum) values are computed as well.

The *gravity axis* is determined according to the highest median value of the sensed gravitational force from the comparison of all axes. Recall that the respective logical sensor is a result of low- and high-pass filters applied to the raw readings from the physical sensors.

Next, the base frequency is computed for computed for each device and each axis. For this, the signals are transformed from the time into the frequency domain using the FFT. The base frequency in Hz is then determined according to the frequency bin with the maximum value.

Note that whereas the former motion-based features are computed individually for each device, additional features are supposed to model the correspondences between both signal streams at a time. For each pair of axes, their *covariance* and *mutual information* are computed. In addition, *acceleration magnitude mutual information* denotes the mutual information of the median magnitudes of the devices' three-axis acceleration measurements, i. e.

$$I(X', Y') = \sum_X \sum_Y p\big[f(\boldsymbol{x}), f(\boldsymbol{y})\big] \cdot \log \frac{p\big[f(\boldsymbol{x}), f(\boldsymbol{y})\big]}{p\big[f(\boldsymbol{x})\big] p\big[f(\boldsymbol{y})\big]} \tag{181}$$

where $f : \mathbb{R}^3 \to \mathbb{R}, \boldsymbol{v} \mapsto \mathrm{Median}(|\boldsymbol{v}_x|, |\boldsymbol{v}_y|, |\boldsymbol{v}_z|)$. At last, *covariance* and *mutual information* are computed for both orientation and rotation rate for each pair of axes at a time.

5.4.1.3 *Audio-based features*

For each audio recording the *base frequency* is calculated individually. The correspondences between the audio recordings of both devices are modeled in terms of their *covariance* and *mutual information.* The *cross-correlation* feature

$$\rho_{X,Y}(\tau) = \mathbf{E}\big[(X_t - \mu_X)(Y_{t+\tau} - \mu_Y)\big] \tag{182}$$

is determined according to $\mathrm{argmax}_\tau\ \rho_{X,Y}(\tau)$.

Next, *audio loudness* and *audio loudness delta* are computed. Audio loudness denotes the median amplitude of the audio signal in the current window. Consequently, audio loudness delta refers to the differences in loudness between both devices.

Post-processing is concluded by computing the MFCCs. MFCCs are widely used for modeling of the perception as in human hearing. The idea is to have a number of coefficients that describe a non-linearly scaled spectrum of a spectrum [287]. This means that first the raw audio signals are transformed into the frequency domain. The frequency components are then squared to determine the signal's power spectrum. Subsequent scaling to non-linear Mel scale helps to describe linear relations between perceived pitch and actual frequency [287]. Dependent on the chosen number M of filter banks (usually between 24 and 40), each of which will correspond to a Mel frequency basis, the spectrum is then scaled according to M triangular windows (e. g. Bartlett), and the logarithms of each window's sum of squares are computed. This relies on the notion that humans usually cannot perceive the differences between closely situated frequencies. Hence the sum of the powers gives an idea of the signal's perceived energy around the M frequencies central to each filter bank. As a consequence of the overlap of the triangular windows, the computed values are correlated. In order to decorrelate them, they are transformed once more, this time using the Discrete Cosine Transform (DCT). The resulting coefficients are the MFCCs. Similar to PCA not all coefficients need to be kept. Coefficients corresponding to regions of lower frequencies usually carry substantially more information than those corresponding to high frequencies. The present system therefore keeps coefficients 0 to 12.

5.5 EVALUATION

Feature vectors were calculated for sliding windows centered around multiples of the selected feature vector rate F_r. The window size w_s was initially chosen equal for all features as $w_s = 1/F_r$, resulting in strictly adjacent and non-overlapping windows. Prior to further analysis of either F_r or w_s, a number of classifiers were compared using 10-fold cross-validation on a dataset corresponding to $F_r = 0.5$ Hz. The results are listed in table 36. Except for Naïve Bayes all of the listed models exhibit good to very good performance. Interestingly enough, Naïve Bayes shows high precision for $C^{\oplus}$, along with acceptable recall for both $C^{\oplus}$ and $C^{\ominus}$, but precision is low for $C^{\ominus}$. Indeed the model classified more than 25% of the instances of $C^{\ominus}$ as $C^{\oplus}$. It is assumed that this is a consequence of Naïve Bayes being the only generative model among the tested classifiers. As a generative model, it is in particular subject to the significant difference between the class priors ($p(C^{\oplus}) = 0.83$, $p(C^{\ominus}) = 0.17$), yielding a strong bias towards $C^{\oplus}$. The remaining models can be considered equivalent in terms of classification performance.

Decision trees were chosen as the default for subsequent evaluations because of their interpretability and the fact that the decision process is easier to comprehend in comparison to the other classifiers for particular samples. Also, the importance of features directly corresponds to their place in the hierarchy of the tree. This may allow social researchers to draw further conclusions about the significance of specific features for related models of real-life social scenarios. In addition to that, parts of a decision tree could also be manually remodeled if necessary. Furthermore, decision trees are deemed as a good fit from a mobile computing perspective. New samples can be evaluated at low costs and model parameters can be adapted without a demand for external processing infrastructure. Although the chosen model is discriminative, it relies on a comparatively low number of model parameters in spite of its high-dimensional input. For decision trees the number of model parameters is a function of several parameters, such as the number and composition of continuous and discrete input variables, as well as possible constraints on the tree itself. In its current form, the decision tree is mostly comprised of binary splits due to mostly continuous random variables, and pruning was performed to get rid of those parts

		$C^{\oplus}$			$C^{\ominus}$		
Classifier	Accuracy	Prec.	Rec.	F_1	Prec.	Rec.	F_1
Naïve Bayes	73.3%	92.5%	73.7%	82.0%	36.6%	71.8%	48.5%
Decision Tree ($J48^{(2)}$)	96.3%	97.5%	98.0%	97.7%	90.3%	88.0%	89.1%
Decision Tree ($J48^{(50)}$)	95.7%	96.4%	98.4%	97.4%	91.7%	82.7%	87.0%
Logistic Regression	95.0%	95.2%	98.9%	97.0%	93.8%	76.4%	84.3%
Neural Network (1HL)	95.8%	96.5%	98.4%	97.5%	91.8%	83.4%	87.4%
SVM	94.4%	94.1%	99.4%	96.7%	96.2%	70.6%	81.4%

Table 36.: Classifier performance for $F_r = 0.5\text{Hz}$ and $w_s = 1/F_r$. The results were computed by 10-fold cross-validation using the Weka toolkit [134]. Note that $J48^{(2)}$ denotes a J48 decision tree with at least 2 samples per leaf whereas $J48^{(50)}$ denotes a minimum of 50 samples per leaf.

of the tree that convey no substantial information. Constraints on the minimal number of samples per leaf had no significant influence on the overall performance, as shown by comparison of the $J48^{(2)}$ and $J48^{(50)}$ decision trees, for which at least two ((2)) and fifty ((50)) samples were required per leaf.

Qualitative inspection of the distributions of the continuous input variables for both $C^{\oplus}$ and $C^{\ominus}$ shows that not all of them are linearly separable. It is therefore suggested that further research should investigate other choices of general model parameters, or possibly feature-specific combinations. Another open question regards the integration of class priors into the model. This is a general issue for discriminative models (see 2.3.4), but not necessarily so for the present model. Cieslak and Chawla describe the construction of trees for unbalanced data [56]. This seems however unnecessary as the high scores for precision and recall for both $C^{\oplus}$ and $C^{\ominus}$ suggest that this model is not biased towards either class in spite of the significant difference between the class priors.

5.5.1 *Feature vector rate and window size*

So far, both feature vector rate F_r and window size w_s were chosen as invariants where $F_r = 0.5$ Hz and $w_s = 1/F_r$. These preliminary values were chosen because computing feature vectors at two second intervals seems to provide an intuitive balance between little and highly varying dynamics in social activities, for which two aspects should be considered: Arguably, social co-activities last for longer periods than just two seconds. Nevertheless it should be possible to detect changes between subsequent co-activities without substantial delay. Recall that precise knowledge of the semantics of the activities not of particular interest for the purpose of co-activity detection. It is however desired that changes between different activities can be detected as such. Without application-specific requirements, two seconds seem to be a reasonable default which will also be verified in what follows. Selecting lower values for F_r may result in higher resolution of the features and thus more information available to the classifier. That way, the distinction of closely related or nested types of co-activities, which would otherwise be difficult to tell apart, may become feasible.

Now in order to assess the selected default for the feature vector rate, a controlled number of variations of F_r were evaluated (see table 37). The corresponding results sustain the default of $F_r = 0.5$ Hz. Beginning with very high performance around $F_r = 8$ Hz, the general trend follows a bathtub-like curve which has its low around $F_r = 0.1$ Hz and then climbs monotonously until $F_r = 0.025$ Hz. Note that w_s was in each case chosen as the reciprocal of F_r, hence ranging from windows of 0.125s to 40s. Also note that the performance for $F_r = 0.5$ Hz is about the same as for $F_r = 0.025$ Hz. The default choice is furthermore ratified by taking into account that

1. the very high performance for $F_r = 8$ Hz may be a result of overfitting, and that
2. the bathtub-like curve illustrates a general trend under which performance decreases at first but then gradually increases to another maximum at the other end of the scale which is merely equivalent in performance.

Next, recall that some of the computed features were based on sensors with a much higher sampling rate than others. In addition to that, some features correspond to random variables whose values are expected to change more often than others. The latter is *per se* irrespective of the sampling rate, although sampling rates are typically chosen proportional to the expected variance. Consider, for example, a sensor such as GPS as opposed to inertial or audiovisual sensors. In regard of the sensor- and feature groups described in section 5.4.1, the intuition that follows from this is that, while F_r is kept constant, for each group of location-, motion- and audio-based features, the window size w_s should be adapted individually according to the *expected change rate* and *expected amount of information* within a given time frame, subject to the following to considerations: Depending on the chosen window sizes, windows of some or all features may eventually overlap. Their size should be chosen so as to avoid overfitting, such as may be the case for very small windows. On a sidenote, one may argue that in order to dampen high-frequency components and to compensate for losses due to non-periodicity of the recorded signals, samples should be scaled using a specific windowing function [359, 304] (see section 5.1.1), especially in the context of overlapping windows. For the current application this is however not necessary as the classifier will only ever consider individual feature vectors, and the presence of high-frequency noise will not have significant impact on the chosen features, e. g. due to the comparatively low resolution of the analysed frequency spectra or due to the use of metrics such as the base frequency.

Several evaluations were performed in order to assess suitable choices for w_s and groups of features. First, window sizes were gradually increased equally for all feature groups from 2s to 10s, 20s, 30s, 45s and 60s (w_s const). Subsequently, two strategies (I) and (II) were evaluated for varying window sizes between the feature groups location, motion and audio (section 5.4.1): Strategy (I) is based on the assumption that window sizes should be chosen “inversely proportional to the average sampling rate of each feature group” [19]. In other words, location-based features should be computed from much bigger windows than

motion-based features than audio-based features. The evaluation of this strategy (I) was performed for choices of $w_s^L = 60s$, $w_s^M = 20s$ and $w_s^A = 10s$ for location, motion and audio, respectively. Strategy (II) is based on the insight that the model may further profit from the availability of additional features, each corresponding to the known features in every feature group, but computed for varying window sizes. This can be explained as follows: Independent of the sensors' sample rates, distinct activities can (and likely will) have completely different profiles with respect to varying window sizes. For example, small windows for an accelerometer's datastream may be useful to distinguish between activities such as dancing and running, but may be completely useless for the discrimination of dancing and playing chess. In terms of window sizes "inversely proportional to the average sensors' sampling rates", smaller increments were chosen for groups with higher sampling rates as opposed to bigger increments for groups with lower sampling rates. Consequently, the evaluation of strategy (II) was performed for choices of $w_s^L \in \{10s, 30s, 60s\}$ and $w_s^M, w_s^A \in \{1s, 5s, 10s, 30s\}$.
Table 38 lists the evaluation results for strategies w_s const, (I) and (II). Note that w_s const differs from the previous evaluation for varying F_r and $w_s = \frac{1}{F_r}$. Instead, F_r is now kept constant at the default 0.5 Hz and w_s is selected independently of F_r. As expected, the current results sustain the prior arguments towards strategy (II). Despite another local maximum of accuracy at $w_s = 10s$ for w_s const, overall both accuracy and F_1 scores reach their optima along with the anticipated information gain through additional windows per feature group. The choice of (II) over (I) and w_s const follows the general trend of the results. What remains is to clarify whether making specific choices for w_s is a way of inserting heuristics into the model, as doing so would possibly contradict the initial postulate for the minimization of heuristics. In regard of varying w_s the extent of this issue is considered negligible. To the contrary, particular and/or limited choices of window sizes are inevitably bound to restrict and influence the model's view of the world. The question is if this can be avoided at all. It is important to note that the present choices have not been made with a particular application in mind. As discussed before, varying window sizes allow the model to learn aspects that could otherwise not be detected, and should thus actually be understood as a way of *generalizing* the model. It should also be mentioned that, while the model for co-activity detection is yet a result of supervised learning, apart from the detection it relaxes the actual recognition of the precise types of activities, thereby allowing the model to gain a much more general understanding of what it means to interact.

5.5.2 *Feature Analysis*

The high performance of the model may lead to further questions regarding potential overfitting and the selection of features [34, 218, 128]. Although the decision tree itself is relatively sparse in terms of parameters, particularly so after pruning, which typically left an average number of 35, 19 and 25 nodes for strategies (I), (II) and ws const, the model is of course still subject to the "curse of dimensionality" that comes along with an

F_r (Hz)		8	4	3	2	1	0.9	0.8	0.7
Accuracy		98.31%	97.56%	97.34%	97.09%	96.43%	96.39%	96.12%	96.10%
Precision	$C^{\oplus}$	98.80%	98.10%	98.00%	97.80%	96.90%	96.90%	97.00%	96.70%
Recall		99.20%	98.90%	98.80%	98.70%	98.80%	98.80%	98.30%	98.70%
F_1 Score		99.00%	98.50%	98.40%	98.25%	97.84%	97.84%	97.65%	97.69%
Precision	$C^{\ominus}$	96.00%	94.80%	94.20%	93.80%	93.90%	93.60%	91.60%	93.10%
Recall		94.30%	91.00%	90.30%	89.30%	85.00%	85.30%	85.70%	84.00%
F_1 Score		95.14%	92.86%	92.21%	91.49%	89.23%	89.26%	88.55%	88.32%

F_r (Hz)		0.6	0.5	0.4	0.3	0.2	0.1	0.05	0.025
Accuracy		96.01%	95.76%	95.64%	95.50%	95.17%	94.66%	95.25%	95.88%
Precision	$C^{\oplus}$	96.70%	96.40%	96.70%	96.70%	96.60%	95.70%	94.60%	95.50%
Recall		98.50%	98.50%	98.10%	97.90%	97.60%	97.90%	100.00%	99.80%
F_1 Score		97.59%	97.44%	97.39%	97.30%	97.10%	96.79%	97.23%	97.60%
Precision	$C^{\ominus}$	92.10%	92.30%	90.30%	89.70%	88.00%	88.80%	100.00%	98.70%
Recall		84.40%	82.70%	84.10%	84.00%	83.70%	79.10%	72.80%	76.50%
F_1 Score		88.08%	87.24%	87.09%	86.76%	85.80%	83.67%	84.26%	86.19%

Table 37.: Performance metrics for $J48^{(50)}$ with varying feature rate F_r, window size $w_s = 1/F_r$. The results were computed by 10-fold cross-validation using the Weka toolkit [134].

$F_r = 0.5$ Hz		w_s const						(I)	(II)
Strategy		2s	10s	20s	30s	45s	60s		
Accuracy		95.76%	97.19%	96.73%	96.77%	96.86%	96.38%	96.30%	97.52%
Precision	$C^{\oplus}$	96.40%	97.30%	97.10%	96.90%	97.50%	97.60%	96.90%	97.70%
Recall		98.50%	99.40%	99.00%	99.20%	98.80%	98.00%	98.60%	99.30%
F_1 Score		97.44%	98.30%	98.00%	98.10%	98.10%	97.80%	97.74%	98.49%
Precision	$C^{\ominus}$	92.30%	96.70%	95.00%	95.80%	93.80%	90.30%	93.00%	96.60%
Recall		82.70%	86.90%	85.80%	85.20%	87.90%	88.80%	85.30%	89.00%
F_1 Score		87.24%	91.50%	90.20%	90.20%	90.70%	89.60%	88.98%	92.64%

Table 38.: $J48^{(50)}$ performance metrics for variations of w_s depending on strategy after 10-fold cross-validation. Strategy (I) corresponds to w_s^L=60s, w_s^M=20s, w_s^A=10s, strategy (II) to $w_s^L \in \{60s, 30s, 10s\}$, $w_s^M \in \{30s, 10s, 5s, 1s\}$, $w_s^A \in \{30s, 10s, 5s, 1s\}$. Superscripts L, M and A denote location, motion and audio, respectively.

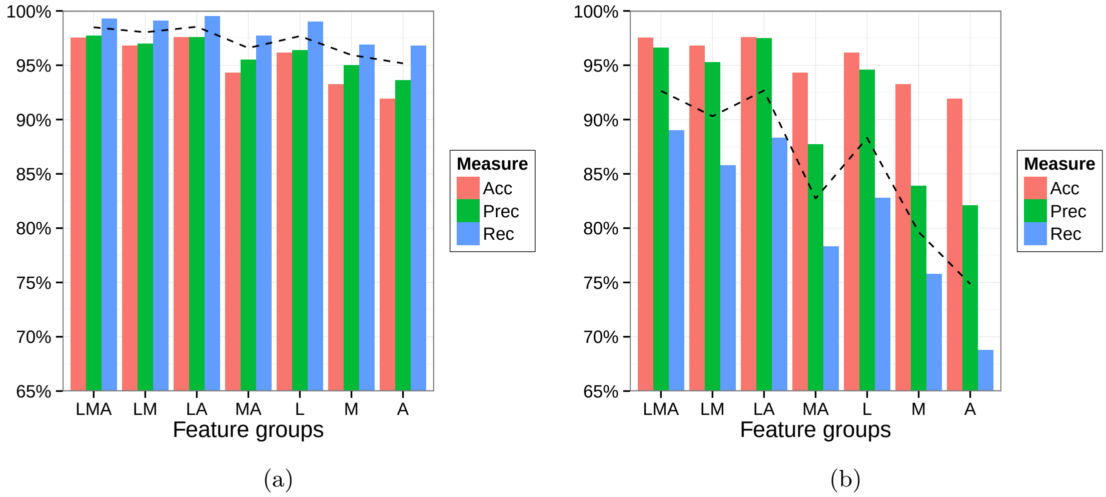


Figure 50.: Ablative analysis of the relevance of the feature groups location (L), motion (M) and audio (A) for co-activity detection. The dashed line denotes the F_1-score.

increasing number of random variables. For the present dataset, the number of features exceeds the number of recorded sessions, which is however alleviated by the length of the sessions and the resolution of the recordings. The aforementioned pruning, in conjunction with the constraint of at least 50 samples per leaf, reduces the number of features that are actually used and hence the number of parameters, and consequently eases the demand for a much greater dataset. The number of parameters will naturally grow the more data become available. Pruning and the constrained number of samples per leaf serve as a more "natural" way of feature selection than other means which are usually applied in order to maximize classifier performance [128] and which may eventually lead to overfitting.

As was mentioned before, the choice of decision trees provides a way to comprehend part or all of the classifier's decision process. In this regard it is interesting to see how those features which were computed for each device separately actually fit into the model. It certainly makes sense that co-activities can be derived from features which take both devices into account. Interestingly enough, the former features still serve their purpose as they tend to come in pairs, yet at different levels within the decision tree. Speaking of levels, it is clear see that the importance of features is directly related to their position in the hierarchy of the tree. This property can just as well be exploited to assess the value of whole groups of features, such as location-, motion- and audio-based features. Visual inspection of the decision trees resulting from the various strategies yields the insight that location-based features are the most important, primarily the proximity feature. This is probably as expected from personal intuition, but it naturally raises the question what will happen if that feature were taken away from the dataset. In order to assess the relevance of the three feature groups, a subsequent analysis was performed during which all of $2^{\{L,M,A\}\setminus\emptyset}$ were evaluated, where L, M, A denote location, motion, and audio. The results in figure 50 corroborate the notion that location-based features seem to be the most important category in co-activity detection, at the very least in terms of the present dataset, followed by motion- and eventually audio-based features. From the results it can also be seen that each feature group on its own contributes to the overall result, since none of the corresponding models fail to produce accurate results. The latter is first and foremost the case for $C^{\oplus}$, whereas for $C^{\ominus}$ one can see that the classifier performance deteriorates from location to motion to audio (as emphasized by the F_1-score as opposed to accuracy). It can therefore be concluded that the model will be susceptible to significantly differing class priors once more and more features were left away. It is also noteworthy that, for both $C^{\oplus}$ and $C^{\ominus}$, LA is en par with LMA, whereas LM shows slightly less performance. In spite of the sole performance of M versus that of A, the latter seems to provide a better supplement in combination with L. This is probably due to the fact that the features in A have less correlation with L than those in M with L, which, apart from the notion that location and motion *may* generally be more closely related than location and audio, is more likely a matter of the nature of the recorded sessions. Further research may investigate the relevance of feature groups with respect to certain kinds or groups of activities. This is however beyond the scope of this thesis.

Apart from whole groups of features it turned out that particular features were much less effective than others. Among the less effective feature were the MFCCs, audio cross-correlation and location speed delta [19]. With respect to the present dataset, removal of any of those features has small to no impact on the model's overall performance. For the MFCCs this comes to no surprise as none of them showed up in any of the decision trees. J48 decision trees are an implementation of the C4.5 algorithm, based on maximization of entropy [254]. Low entropy may hint at the lack of speech or characteristic noise, both of which may go hand in hand. As the preferred on-body location is the front-pocket of the trousers [150], it can be assumed that MFCCs indeed are among the more irrelevant features, neglecting the fact that microphones were worn openly during the recording of this dataset. On the other hand, there may be situations (activities) which are much more characterized by speech, or where the phone is e. g. placed on a surface, so that the MFCCs may yet contribute to the process. The desire for a preferably universal model, however, plus the fact that the inclusion of the MFCCs in the present evaluation had no negative consequences, suggests that they should not be left out of further considerations. As far as audio cross-correlation and location speed delta are concerned, their absence in the decision tree is very likely also due to the limitations of the present dataset. A good example for when location speed delta could be relevant is given in [19], where it may help to discern fast-paced activities, e. g. due to sports or transportation, from other kinds of activities. In a related sense this applies to audio cross-correlation as well, when for example loud or very characteristic (e. g. rhythmic) environments ought to be differentiated from others.

The last question in this section is concerned with the permanent or temporal lack of certain features due to the outage of physical or logical sensors. A naïve solution would be the computation of separate models for each case of singular or combinations of multiple missing features. This is clear intractable as it would require the computation of up to 2^{N-1} models, given a set of N features. It may however be feasible to determine groups of features which are likely to fail together, such as all features that rely on e. g. the presence of accelerometer or gyroscope readings, which would greatly reduce that overhead of the number of "spare" models. As a last resort, a whole feature group such as L, M or A, could be left out, resulting in the previously seen seven distinct models (refer to figure 50). An alternative solution could be to stop the evaluation of the decision tree for a particular sample at that very node v for which the corresponding feature could not be computed or is simply missing. In that case it seems reasonable to perform a majority voting according to distributions of $C^{\oplus}$ and $C^{\ominus}$ at the leaves of the subtrees of v. There is however a high risk of leaving out features that are actually present somewhere deeper in the hierarchy, which is why it is suggested to stick with the former approach of an intelligent choice of "spare" models.

5.6 CO-ACTIVITY SEGMENTATION AND CLUSTERING

Applications in SSP could be interested in more than just the simple fact that two individuals were performing the same co-located activities during a given period of time. It may for instance be interesting to know whether one or more distinct activities were performed during that time. Even though the proposed concept does not require the precise types of these activities to be known, a number of social aspects may be derived from this information, for which examples were given in the previous sections. In addition to changes in the activity types, applications may furthermore be interested in recognizing equal activities that were not performed in timely sequence. Altogether this implies a demand for segmentation and subsequent clustering of a stream of co-activities. As the precise activity types will obviously not be known in real world settings, both segmentation and clustering need to be implemented in an unsupervised fashion.

As part of the evaluation of the framework that was developed during the proceedings of this thesis [19], Bader used an EM-based clusterer, provided by the Weka toolkit [134], which he applied to those feature vectors that were previously positively detected as co-activities for each session. The clustering algorithm is based on multivariate GMMs and iteratively adds new clusters until there is no further increase in log-likelihood after 10-fold cross-validation. Each detected cluster corresponds to a single activity type. Ordering the feature vectors by the exact times which they represent consequently leads to a timely sequence of detected activity types. This sequence is then smoothed by using a moving median to compensate for outliers, which is further justified by the assumption that activity types will not change back and forth at sub-second intervals. Unfortunately, the evaluation in [19] is flawed because the author did not account for the fact that after positive identification of co-activities in a session, these co-activities are not necessarily adjacent. As a consequence, changing points are identified in a presumed sequence of co-activities which actually contains gaps. Instead, sessions should have been split into subsessions of continuous co-activity prior to evaluation.

Generally speaking, EM-based approaches are advantageous in situations where the number of clusters is *a priori* unknown. In a scenario such as co-activity segmentation, however, it is very likely that the number of detected clusters will differ from the ground-truth number of activity types. One important fact in this matter is the already discussed recursive nature of activities. Naturally, new clusters show up along with significant changes in the distribution of the measured variables, whereas those changes do not necessarily imply an actual change in the activity type as perceived by human experts who label the data. On the other hand, the use of GMMs in EM-based clustering makes the process more robust against outliers, helps to compensate for missing data, and allows for clusters of different size and correlation between the variables (e. g. as opposed to K-Means). Arguably, the disadvantages of EM-based approaches are their computational complexity, a usually sensitive choice of constraints for the covariance matrices of the Gaussians in order to prevent singularities, especially when facing high-dimensional data, as well as the fact that all data are taken into account at once. This global view may for instance yield a clustering which may be optimal in terms of likelihood, but basically miss out on clusters which otherwise could have been detected from a *local* perspective. More precisely, by taking into account all samples at once, potential implications of the timely sequence of the samples are lost. For example, from a number of samples, all of which actually belong to the same cluster, a portion may be associated with another overlapping cluster when seen from a global rather than a local point of view, a fact otherwise naturally justified by the i.i.d. assumption. The EM-based approach is also not well suited for processing streams of activities. First, the stream cannot easily be split into chunks that could then be processed by EM as a whole because that may lead to overlapping segments of activities, especially in cases where changes would occur close to segment borders. Second, even though EM-based approaches can be adapted to online variants that can be fed additional data, this would come at the price of losing robustness and the search for a suitable stopping criterion.

As a consequence, the work at hand proposes a two-fold process in which co-activities are first segmented in a top down approach and then clustered from the bottom up. This choice is motivated by corresponding techniques for *speaker diarization*, also known as speaker segmentation and clustering [324]. Note that speaker diarization systems usually involve decoding steps that separate speech from non-speech segments in advance of further processing. The proposed system is equivalent in this sense since it separates co-activities from non-co-activities before segmentation and clustering, and can therefore be considered as a *co-activity diarization system.*

5.6.1 *BIC-based Segmentation*

Segmentation systems are typically categorized as decoder-based, model-based or metric-based [53, 175]. Decoder-based systems are only concerned with separating speech from non-speech at points of silence or, respectively, co-activity from non-co-activity at points where there would be no measurable activity at all. Model-based segmentation, on the other hand, relies on a fixed set of models according to an *a priori* selection of specific classes such as speech, music, or noise, but also individual speakers, which would relate to a specific choice of previously selected co-activities. Finally, the last category of systems is based on finding the extrema of chosen metrics between moving adjacent windows.

Such metric-based approaches are generally considered to yield high recall at moderate precision [175]. An example of a corresponding metric is given by the KL2 distance metric as introduced by Siegler et al. [302], which for two distributions A and B is defined as the (symmetric) sum of the (asymmetric) KL divergences from A to B and B to A, supposedly yielding better results than previous model-based approaches. The most notable metric used in segmentation systems is the BIC-based generalized likelihood ratio [53, 324, 175]. The idea is to regard the input stream as a Gaussian process and identify those changing points t for which it holds that the process is best modeled with two distributions instead of just one if it were split at t. Readers should be aware that in contrast to systems which also consider overlapping speech [39] the activity type during co-activity is by definition unique.

In their seminal work on the BIC criterion for segmentation, Chen et al. propose testing the hypothesis H_0 that a change occurs at time t versus the alternative H_1 that there is no change, and therefore whether the data around t should be modeled with two rather than a single distribution [53]. For this, let $\mathbf{X} = \{x_1, \ldots, x_N\}$ be a set of multivariate samples. For a given changing point $1 < t < N$, the likelihoods $\mathcal{L}_1$ of a model M_1 with parameter set θ_1^1, and $\mathcal{L}_2$ of a model M_2 with parameter sets θ_1^2, θ_2^2 are then given by

$$\mathcal{L}_1 = \prod_{i=1}^{N} p(x_i|\theta_1^1) \quad \text{and} \quad \mathcal{L}_2 = \prod_{i=1}^{t} p(x_i|\theta_1^2) \prod_{i=t+1}^{N} p(x_i|\theta_2^2) \,. \tag{183}$$

The maximum log-likelihood statistic for H_0 and H_1 consequently is

$$\begin{aligned} \log \frac{\mathcal{L}_2}{\mathcal{L}_1} &= \log \mathcal{L}_2 - \log \mathcal{L}_1 \\ &= \sum_{i=1}^{t} \log p(x_i|\theta_1^2) + \sum_{i=t+1}^{N} \log p(x_i|\theta_2^2) - \sum_{i=1}^{N} \log p(x_i|\theta_1^1) \,. \end{aligned} \tag{184}$$

Then, since M_2 has twice as many parameters as M_1, the difference of their BIC values is

$$\Delta \mathrm{BIC}(t) = \sum_{i=1}^{t} \log p(x_i|\theta_1^2) + \sum_{i=t+1}^{N} \log p(x_i|\theta_2^2) - \sum_{i=1}^{N} \log p(x_i|\theta_1^1) - \lambda \frac{k}{2} \log N \tag{185}$$

for $\lambda = 1$ and k the number of parameters according to any one of θ_1^1, θ_1^2 or θ_2^2. It is furthermore clear that a model which has more parameters yields at least equal or better likelihood for the same data. Thus H_0 must be rejected for any given t if $\Delta\mathrm{BIC}(t) \leqslant 0$ because then the gain in likelihood would not outweigh the penalty induced through the additional parameters. As the maximum likelihood estimate of a changing point is given by $\hat{t} = \mathrm{argmax}_t \, \Delta\mathrm{BIC}(t)$, it follows that any such candidate changing point $\hat{t}$ will be considered as a true changing point if and only if $\Delta\mathrm{BIC}(\hat{t}) > 0$.

Figure 51 shows the result of computing $\Delta\mathrm{BIC}(t)$ for varying t on data from two adjacent segments of distinct co-activities. From the figure one can clearly see that the peak of the metric coincides with the true changing point. It should also be noted that $\Delta\mathrm{BIC}(t)$ was

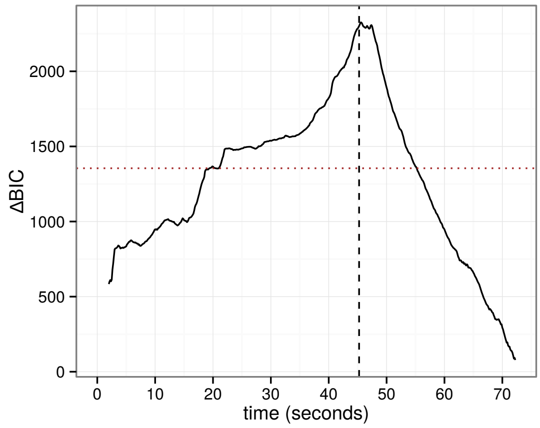


Figure 51.: The ΔBIC metric for two adjacent segments of distinct co-activities. The dashed line denotes the true changing points.

only computed for values of t which lie several seconds apart from the session borders. The reason for this is three-fold: First, enough samples are needed so that covariance does not collapse onto a single point. Second, the number of samples must well exceed the number of model parameters to achieve useful measures of likelihood. Third, if the window size were chosen to large, the window might actually contain more than a single changing point, thus violating the model assumption [358]. A co-activity model which is likely based on tens or hundreds of features will therefore require sufficiently large windows or a systematic increase of the feature vector rate F_r. The comparison of varying window sizes for the present dataset has shown that $w_s = 10s$ is a reasonable value for segmentation. On other hand, according to the prior evaluation of the model not all features carry significant contributions to the overall problem of predicting $C^{\oplus}$ versus $C^{\ominus}$. As a consequence, another means of avoiding large window sizes or increased sampling rates is given by means of information reduction, e. g. by application of PCA. In fact, the following evaluation will show that a projection of the data onto a relatively small number is well sufficient for the task.

The problem of choosing sufficiently large windows is likewise known in speaker diarization, which is why Chen et al. [53] defined *detectability* D as a measure of how likely it is to detect a true change as $D(t) = min(t, N - t)$ for given t and window size N. This is an important insight as it implies that changes will be hard to detect whenever they lie close to the window's borders. Aside from these drawbacks, the probabilistic roots of the metric have clear advantages. For one, it can be shown that along with an increasing number of samples the maximum likelihood estimate $\hat{t}$ converges against a true changing point. Furthermore, since ΔBIC takes into account all samples from a window at once, the metric is potentially much more robust than metrics that operate separately on each split, such as KL2. A manually chosen threshold is also not necessary. Instead, the penalty term of BIC provides an implicitly determined threshold. Manual fine-tuning is still possible through *a posteriori* adaption of λ in equation (185), a fact naturally exploited in related work [326, 324]. Moreover, λ could also be chosen dependent on $\hat{t}$ to compensate for lower likelihoods due to fewer samples in smaller windows at segment borders.

The principle of maximizing $\Delta\mathrm{BIC}$ has yet to be implemented as a suitable algorithm for the current task of co-activity segmentation. Calculating $\Delta\mathrm{BIC}(t)$ for all $1 < t < N$ of a segment of N samples and subsequently determining its peak is not applicable in the general case in order to avoid violating the model assumption whenever the data contain more than one true changing point. Under this consideration, the original algorithm as proposed in [53] is outlined as follows:

```
$\mathcal{C} = \{\}$                                  ▷ a set of detected changing points
$w_0 \leftarrow 10s$                   ▷ a reasonably chosen initial window size, e. g. 10s
$a \leftarrow 0$                              ▷ the left boundary of the current window
$b \leftarrow a + w_0$                       ▷ the right boundary of the current window
while b < N do
    $\hat{t} \leftarrow \mathrm{argmax}_{a<t<b}\ \Delta\mathrm{BIC}(t)$
    if $\Delta\mathrm{BIC}(\hat{t}) > 0$ then
        $\mathcal{C} \leftarrow \mathcal{C} \cup \{\hat{t}\}$          ▷ add $\hat{t}$ to the set of detected changing points
        $a \leftarrow b + 1$
        $b \leftarrow b + w_0$
    else
        $b \leftarrow b + 1$
    end if
end while
```

Related work suggests a number of improvements, such as variable window schemes or avoiding computations of $\Delta\mathrm{BIC}(t)$ where detectability $D(t)$ is below a chosen threshold [326]. Furthermore, if the metric has its peak near the end of a segment, then it is likely that a true changing point approaches. Therefore, the window size should not be increased beyond reason for the next step, as that may lead to oversized windows in scenarios where the data contain frequent changes [358]. The algorithm was initially proposed for use with single multivariate Gaussians, but other models (e. g. GMMs) were also reported in subsequent works [53, 26, 210, 267, 324, 175]. For the case of single Gaussians, Zhou and Hansen [358] propose using Hotelling's two-sample T^2 statistic to avoid redundant computations of the determinants of the covariance matrices, yielding significant speedups in regard of the algorithm's quadratic complexity. The T^2 distribution generalizes Student's t distribution [145] and can be used to determine the likelihood of different means for multivariate distributions. For a segment $\mathbf{X}$ of N samples and a candidate changing point $\hat{t}$, the T^2 statistic is defined as

$$T^2 = \frac{t(N-t)}{N}(\mu_1 - \mu_2)^T \Sigma^{-1} (\mu_1 - \mu_2) \tag{186}$$

where μ_1, μ_2 denote the means of each subsegment (window), and Σ denotes the covariance matrix of $\mathbf{X}$. Most notably, it can be shown that the peak of this T^2 statistic corresponds to the maximum likelihood estimate $\hat{t}$. The T^2 statistic can therefore be used as a preliminary step to estimate $\hat{t}$, so that $\Delta\mathrm{BIC}(\hat{t}) > 0$ has to be evaluated only once.

The proposed algorithm for the segmentation of co-activities is closely related to the outlined algorithm. As input the algorithm expects a sequence of contiguous feature vectors, previously classified as $C^{\oplus}$. Based on a minimum window size of $w_{min} = 10s$, chosen in order to ensure that windows will contain at least a reasonable number of samples for likelihood estimation, any input sequence of less than 20s will be returned as a single segment. Longer sequences will be processed by starting from an initial window size $w_0 = 2w_{min}$, increased by 5% at every iteration until either a changing point is found or w exceeds the size of the segment. Other than previously outlined though, the BIC test is further constrained. Instead of testing the maximum likelihood estimate $\Delta BIC(\hat{t})$ for positive values only, the test actually requires that

$$\Delta BIC(\hat{t}) \geqslant \text{Median}\big(\{\Delta BIC(t) \mid a < n \cdot t < b,\ n \in \mathbb{N}_0\}\big)\ , \tag{187}$$

so that changing points will be characterized by distinctly attenuated peaks. Irrespective of F_r, computations of $\Delta BIC(t)$ are performed at integer multiples of 1s. Models are based on single multivariate Gaussians because the use of other models such as GMMs entails no significant improvements with respect to the present dataset as opposed to a substantial increase in computational complexity. Note that the proposed algorithm does not involve any prior or posterior adaptions of λ to the dataset as that would otherwise imply a loss of generality.

5.6.2 *Clustering*

Once a session has been split into segments of distinct activity types, which can actually as well be understood as a clustering task in its own right, *non-adjacent* segments of the same activity type should be recognizable as belonging to the same activity. Therefore clustering must be performed for all non-adjacent segments originating from a single stream of co-activities. Same as with segmentation, the clustering process needs to be unsupervised instead of relying on previously learned models for specific classes, as the identity of the recorded activities is unknown [302]. A number of deterministic or probabilistic approaches have been successfully used in speaker diarization [175]. The predominant approach [26, 324] however is based on the same BIC metric as the segmentation algorithm proposed in section 5.6.1.

From the previous discussion it is known that a candidate changing point $\hat{t}$ is considered as a true changing point if and only if $\Delta BIC(\hat{t}) > 0$, since then the data are best modeled with two distributions instead of just one, considering a gain in likelihood that exceeds the penalty introduced by doubling the number of model parameters. Naturally, this criterion implies the opposite in case of $\Delta BIC \leqslant 0$, in other words that two separate clusters should be joined into one. Barras et al. [26] describe the standard BIC based clustering algorithm as follows: Starting from a set of $\mathcal{S} = \{s_1, \ldots, S_N\}$ segments, for each pair (s_i, s_j) with $i \neq j$ compute ΔBIC_{ij} for a model M_1 comprised of a single Gaussian for all samples in $s_i \cup s_j$, and a model M_2 of separate Gaussians for each of s_i and s_j. If $\Delta BIC_{ij} < 0$ for $\text{argmin}_{ij} \Delta BIC_{ij}$, then join the segments s_i and s_j into a single cluster. The process is repeated until $\Delta \mathrm{BIC}_{ij} > 0 \; \forall i, j$.

As a consequence of the fact that adjacent segments were just split based on the very same criterion, only non-adjacent segments need to be taken into account. Along with application-specific choices for λ in equation (185), further parameters may be integrated e. g. for *a posteriori* fine-tuning of segment versus cluster penalties [302, 26]. The maximum likelihood estimate naturally leads to a suitable stopping criterion once no further increase in likelihood is expected, thereby implicitly leading to an automatic estimation of the number of clusters [175, 210]. Moreover, the proposed bottom up approach yields a certain prospect of *locality*. As opposed to other algorithms which take into account all data at once, this approach allows for focussing on data which may be located close-by on (small) temporal scales. In fact, this notion of locality has been shown to be advantageous for clustering of segments of speech as opposed to global processing of the data [26]. On a sidenote, the proposed approach is also suitable for online processing, although it is not considered optimal since an online clusterer tends to converge into local rather than global maxima [326, 324].

5.6.3 *Evaluation*

5.6.3.1 *Evaluation of BIC-based Segmentation*

Evaluation of the proposed algorithm was performed on data for which $F_r = 8\mathrm{Hz}$ and $w_s = \frac{1}{F_r}$. After prior classification only those data corresponding to $C^{\oplus}$ were kept. The dataset was then partitioned according to the 33 annotated sessions. In order to account for non-contiguous sequences due to back and forth transitions between $C^{\oplus}$ and $C^{\ominus}$ in the actual sessions, each session was further split into contiguous sequences of $C^{\oplus}$, yielding a total of 39 subsessions. As the algorithm is based on models of single multivariate Gaussians, PCA was performed to maximize variance and avoid singular covariance matrices, such that the resulting components would account for at least 95% of the original variance. For this, missing values were replaced by the mean for numeric and the mode for categorical variables. Recall that the values of a categorical variables each correspond to a specific state. Also recall that, whereas nominal values could easily be mapped to integer values, PCA takes into account the correlations of the remaining variables with any specific state, instead of some arbitrary magnitude of an integer value of a random variable corresponding to several states. Therefore, each categorical variable of a set of K distinct values was replaced by K boolean variables for a 1-of-K binary encoding.

Figure 52 illustrates a projection of the data onto the first three principal components. One can see that the decorrelation of the variables lead to reasonable separability, particularly so for throwing/catching, walking and jogging, albeit much less for sitting, eating and standing. It is not surprising that throwing/catching and walking interfere with each other since the former is also comprised of phases where subjects move or walk. Sitting, eating and walking seem hardly separable, although that is somewhat mitigated by the third component, and inspection of the remaining components shows additional contribu-

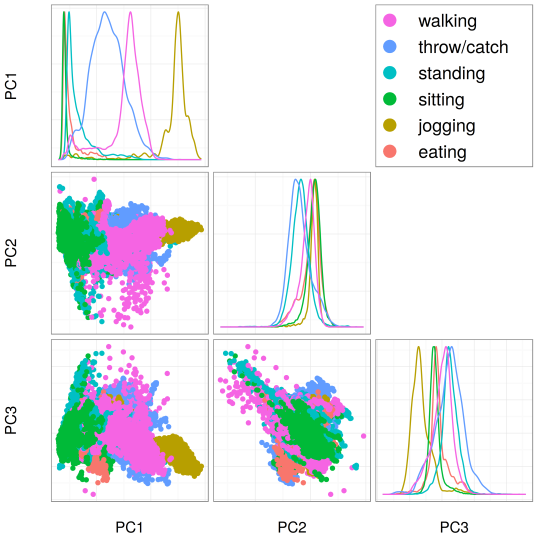
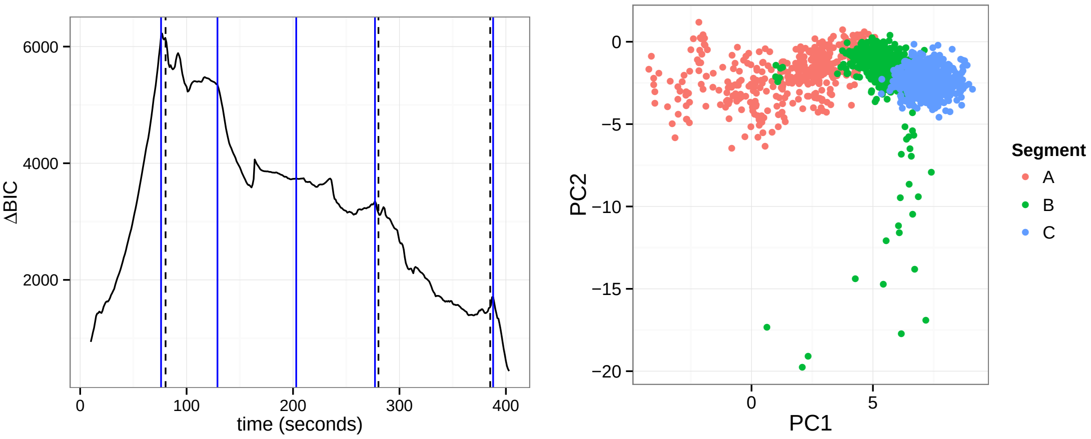


Figure 52.: Distribution of activity types after projection onto the three major principal components.

tions to their separability.

The major portion of the first five principal components is comprised of the cross-correlation of the audio signals of both devices, individual means of acceleration, as well as the mutual information of parts of the orientation quaternions of both devices. This is somewhat expected as these features also often appeared near the root of the decision trees during the discrimination of $C^{\oplus}$ and $C^{\ominus}$ (refer to section 5.5). Interestingly though, summing up over all principal components exhibits the attenuation of audio features such as the MFCCs, which, to the contrary, did not contribute to the decision trees. This can however be explained by the fact that the J48 algorithm, which was previously used for building the decision trees, aims at the maximization of Shannon entropy whereas PCA simply maximizes variance. In the context of the proposed segmentation algorithm, the latter is justified since the segmentation process is primarily governed by the variance of the data.

Application of the proposed segmentation algorithm to a sequence of contiguous co-activities leads to results like the one illustrated in figure 53a. The algorithm has clearly identified the three changing points from ground truth (as shown by dashed lines), yet it has furthermore identified two additional changing points, effectively partitioning a single annotated activity into three potentially different activities. Note that the segmentation does not imply anything about the relation of the second and fourth segments, and hence the correspondingly performed activities. Instead, it merely yields information about differently distributed data from the second to the third as well as from the third to the fourth segment, so that the second and fourth segment might still conform to the same type of activity. From figure 53b one can see that the data indeed follow different distributions, and that activity B furthermore contains a number of outliers. In fact, listening to the recorded audio shows that during these segments, the two recorded subjects first



(a)

(b)

Figure 53.: Exemplary results after automatic segmentation of a sequence of contiguous co-activities (a), along with a visualization of the data's distribution for the second (A), third (B) and fourth (C) segment. The dashed lines denote true changing points, whereas the blue lines correspond to detected changing points.

left their appartment, descended a flight of stairs, the latter of which happened to be in a rather reverberant environment, and then went for a walk surrounded by considerable traffic. Activity A was also characterized by a disproportion between the first and second speaker, which then turned around during activity B. Activity C, although visually not much different from activity B, was furthermore characterized by loud noise of a car driving by. Other than that, the timely discrepancies between the annotated and the automatically detected changing points are acceptable. Generally speaking, a "smooth" transition between subsequent activities is expected, as is a result of human behaviour for which strict and abrupt changes are presumably the exceptional case. For the same reason, manually annotated data likely yield uncertainty. In case of the present dataset, the data were not annotated during but after the recordings, so as to avoid obtrusive sensing to a degree where persons would change their behaviour (see section 1.1.6.1).

All in all, the segmentation algorithm split 34 out of 39 subsessions into 190 segments. The remaining 5 subsessions were too small to be split (shorter than the required $2w_{min} = 20s$.). Aside from precision and recall, the performance of speaker segmentation systems is likewise assessed by their False Alarm Rate (FAR) and Missed Detection Rate (MDR) [175], defined as

$$\mathsf{FAR} = \frac{\mathsf{FA}}{\mathsf{GT} + \mathsf{FA}} \quad \text{and} \quad \mathsf{MDR} = \frac{\mathsf{MD}}{\mathsf{GT}} \tag{188}$$

for FA the number of false alarms, MD the number of missed detections, and GT the number of true changing points. This further leads to the question for reasonable sensibility of a segmentation system. From figure 53a one could already see that a certain discrepancy between the detected and the annotated changing points is likely. Following the prior discussion, evaluation was performed for sensibilities of either 15s or 30s. Precision, recall, FAR and MDR were furthermore computed for varying numbers of principal components in order to find out how much information (in terms of variance) is actually needed for the task at hand. The results are shown in tables 39a and 39b. In this setup, the performance of the system is clearly far from acceptable. Further analysis shows this to be the result of two major issues: For one, the segmentation algorithm has found more changing points than actually present from the annotation. Not unexpectedly, though, consequent inspection of the sample distributions sustains the respective decisions of the system, likewise discussed above in regard of the example from figure 53a. Closer inspection furthermore shows that a non-negligible number of changing points is located close to the borders of the corresponding sessions. It was discussed that such changing points yield only poor detectability [53]. In fact it turned out that these changing points mostly correspond to the beginning of the recorded sessions during which the subjects would often briefly stand and discuss the session, upon which they would typically begin with their first actual activity as laid out in the session scriptlets (see section 5.4). Yet others correspond to uncertainties in the annotation, e. g. to a transition between activities near the ending of a contiguous sequence of co-activities at which one subject for instance left the scene. Therefore the system was evaluated once more, this time disregarding those changing points which lie beyond a certain margin from the sessions' borders. The results in tables 39c and 39d show that respecting a correspondingly chosen margin leads to significant improvements in recall and MDR, whereas the change had no apparent influence on FAR and precision.

As indicated above, the high FAR may be the result of diverging distributions of the samples within segments which were actually annotated as a single type of activity. On the one hand, this may lead to a demand for a more detailed view and consequent annotation of the data. On the other hand, however, this goes along with the different ways and levels at which humans perceive the affective meaning of any activity. It has already been discussed that activities can be nested recursively, depending on the point of view and more precisely also the chosen spatio-temporal frame. Walking, for instance, can be understood as a superposition of smaller activities such as lifting a foot or moving a limb, and in the same sense it could be broken apart into any of these fractions. Furthermore, so far the model has no way of telling apart e. g. the sudden appearance of loud noise from a real change in activities. To the best of the author's knowledge, research has not yet defined a common baseline for this problem. Other than recall and MDR, precision and FAR therefore seem to be inadequate measures for the present task. Nevertheless, lack thereof may be mitigated by analyzing the prevalent activities within each of the automatically determined segments. Recall that co-activity diarization is primarily concerned with finding non-adjacent segments of equal activities, yet – arguably – not necessarily the precise changing points as such. It is thus most important that segments of actually different activities do not overlap. In other words, automatically determined segments should not correspond to more than one distinct type of activity. Unless taken to the extreme, for instance when a session of N samples were split into N segments, this measure can at least assist in the verification of the approach. Table 40 therefore shows the percentage of automatically determined segments corresponding to only a single type of activity. For

# Comp.	FAR	MDR	Prec.	Rec.
5	0.510	0.613	0.271	0.387
10	0.617	0.613	0.193	0.387
15	0.619	0.600	0.197	0.400
20	0.615	0.600	0.200	0.400
30	0.623	0.640	0.179	0.360
40	0.597	0.613	0.207	0.387

(a) Performance of the segmentation algorithm for varying numbers of components. Sensibility 15s.

# Comp.	FAR	MDR	Prec.	Rec.
5	0.490	0.507	0.339	0.493
10	0.603	0.493	0.250	0.507
15	0.605	0.480	0.253	0.520
20	0.603	0.493	0.250	0.507
30	0.601	0.480	0.257	0.520
40	0.588	0.547	0.241	0.453

(b) Performance of the segmentation algorithm for varying numbers of components. Sensibility 30s.

# Comp.	FAR	MDR	Prec.	Rec.
5	0.607	0.409	0.277	0.591
10	0.712	0.432	0.187	0.568
15	0.712	0.409	0.193	0.591
20	0.709	0.409	0.195	0.591
30	0.720	0.455	0.175	0.545
40	0.699	0.432	0.197	0.568

(c) Performance of the segmentation algorithm for varying numbers of components. Sensibility 15s. Margin 30s.

# Comp.	FAR	MDR	Prec.	Rec.
5	0.593	0.318	0.319	0.682
10	0.703	0.318	0.224	0.682
15	0.703	0.295	0.230	0.705
20	0.701	0.318	0.226	0.682
30	0.705	0.295	0.228	0.705
40	0.690	0.341	0.228	0.659

(d) Performance of the segmentation algorithm for varying numbers of components. Sensibility 30s. Margin 30s.

Table 39.: Performance characteristics of the segmentation algorithm.

	Threshold		
# Components	99%	95%	90%
5	68.7	77.1%	81.3%
10	75.1	79.9%	82.8%
15	75.4	79.1%	83.4%
20	75.1	78.9%	83.3%
30	75.2	79.0%	82.4%
40	74.4	78.4%	82.4%

Table 40.: Percentage of segments corresponding to a single type of activity, i. e. those segments for which the number of samples for a single type of activity exceeds the given threshold.

this, a segment corresponds to a single type of activity whenever the number of samples for that type of activity exceeds a chosen threshold, such as e. g. 95%. Together with the prior measures for recall and MDR, these results show that the system runs with acceptable performance, which may yet not be convincing but is at least significantly better than chance.

5.6.3.2 *Evaluation of BIC-based Clustering*

The proposed clustering algorithm was evaluated for each recorded session on both actual as well as ideal results from prior segmentation. The latter are directly inferred from the annotated ground-truth. The reason for this is rooted in the fact that segmentation errors will eventually lead to clustering errors. Clearly, whenever segment boundaries are not properly detected, data from one segment will leak into the other [326, 175]. In the context of co-activity diarization this is mitigated by presuming "smooth" transitions instead of abrupt changes between subsequent activities. This presumption is further corroborated by inspection of the actual data around segment borders, which reveals that in most cases the corresponding samples move gradually from one distribution to the next. This goes hand in hand with the prior results from table 40, from which it follows that segment borders may not be detected precisely where annotated, yet still more than 90% respective 95% of the data correspond to a single type of activity. At the bottom line, the second evaluation should yield a better measure for the clustering process itself, whereas evaluation based on the actual segmentation results should give a better measure for the framework as a whole.

Table 41 shows the results. This time λ was manually optimized for the process. The necessary adaption of λ is a consequence of the reduced size of the segments after segmentation. For this, recall that in order to compensate for the penalty term in equation (185), likelihoods have to be computed for a considerable number of samples. As a rule of thumb, λ was chosen such that $\lambda_c = -1.5 \cdot \frac{40}{c}$, where c denotes the number of components. Next to λ, table 41 shows the number of segments *after* clustering in comparison to their numbers *before*. Clearly, the number of segments before clustering is constant for the ideal scenario whereas it varies strongly in case of the actual segmentation. The table also shows the number of non-adjacent segments that were joined in clusters, for which the respective percentages simply correspond to the fraction by which the total number of clusters were reduced. Although these fractions are given with respect to the total number of segments instead of the number of non-adjacent segments before clustering, they serve to show that the clusterer operates in roughly the same range, irrespective of the ideal or actual scenario. Following the prior discussion from section 5.6.3.1, the ratio of segments for which the number of samples for the prevalent activity exceeds the given threshold, together with the relation between segments before and after clustering, indicates good overall performance of the proposed algorithm. Comparing the results of the ideal scenario to those in table 40 furthermore shows that the subsequent clustering step can improve the performance of the overall diarization process, as it reduces the intra-segment dispro-

Scenario	# Components	λ	# Segm. (of #)	# Clustered (%)	Threshold	Ratio
Ideal	5	-12	107 (134)	27 (20.2%)	0.90	97.20%
					0.95	96.26%
					0.99	92.52%
	15	-3	116 (134)	18 (13.4%)	0.90	97.41%
					0.95	96.55%
					0.99	93.97%
	40	-1.5	114 (134)	20 (14.9%)	0.90	96.49%
					0.95	95.61%
					0.99	93.97%
Actual	5	-12	132 (166)	34 (20.5%)	0.90	78.03%
					0.95	73.48%
					0.99	64.39%
	15	-3	173 (211)	38 (18.0%)	0.90	80.35%
					0.95	76.88%
					0.99	71.68%
	40	-1.5	176 (199)	23 (11.6%)	0.90	80.68%
					0.95	76.70%
					0.99	72.16%

Table 41.: Performance of the clustering algorithm based on actual and ideal prior segmentation, showing the number of segments after vs. before clustering, the number and fraction of clustered non-adjacent segments, and the ratio of segments for which the internal count of the prevalent activity type exceeds the given threshold.

portion between the number of samples of those activities which leaked into the segment due to segmentation errors, and the number of samples of the prevalent activitiy.

6 CONCLUSION AND FUTURE WORK

This work has investigated new means of modeling, capturing and characterizing social context on small spatio-temporal scales through the use of mobile agents without dependencies on external infrastructure. It was discussed that social relationships constitute an elementary aspect of the social context. They are quantifiable as functions of social interaction which can be inferred from social signals and behavioural cues as part of non-verbal communication. It was shown how behavioural cues from social interaction geometry can be used to infer social situations, defined as co-located face-to-face social interaction subject to full mutual awareness of all participants. Once detected, a social situation is described by a four-tuple $S = (P, T, X, K)$ for a set P of persons, a temporal reference T, a spatial reference X, and K a set of tags which may be used to describe the semantics of the situation. Interaction geometry models spatio-orientational arrangements in terms of pairwise measurements $(\delta\theta, \delta\varphi, \delta d)_{ij}$ for persons i and j (as seen by i), where $\delta\theta$ denotes the angle between the shoulder-lines, and the polar angle $\delta\varphi$ as well as the interpersonal distance δd correspond to the relative position. It was shown how a quantitative model based on these dyadic measurements can be used to algorithmically infer whether i and j do or do not interact, and how interaction in groups of $N \geqslant 2$ persons allows for the determination of social situations as a whole. For this, a new dataset was recorded using a high-performance infrared tracking system, and the data were annotated according to the presence ($S^{\oplus}$) or absence ($S^{\ominus}$) of social interaction for each pair of subjects and point in time ($F_s = 6\text{Hz}$). Analysis of the dataset led to the use of mixture distributions to model the experimental data as they employ the means for probabilistic soft-clustering, allow for modeling clusters of varying size and shape, and foster the easy integration of class priors. A new algorithmical model for the detection of social interaction was introduced which discriminates between $S^{\oplus}$ and $S^{\ominus}$ based on separate models for observations from either class. The proposed model is human-interpretable and allows for insight into the decision process, in particular also for researchers from socio-psychological fields. Based on quantitative data the model's decision process makes no further assumptions about specific arrangements such as circular formations [13, 67]. It has been shown to be universally applicable to groups of varying size and in various formations.

It was discussed that a model of interaction geometry should respect the fact that $\delta\theta$ and $\delta\varphi$ are both 2π-periodic. SW-GMMs permit the integration of both linear and non-linear random variables by wrapping them according to their periodicity. As EM-based training requires an increased number of modes and due to the fact that N variables with K tilings into every direction cause a computational overhead of $(2K+1)^N$, it is necessary to restrict K for which it was shown that 3 wraps for each of $\delta\theta$ and $\delta\varphi$ yield accurate descriptions of the dataset. Unexpected at first, the evaluation revealed that the more

correct, yet also more complex, SW-GMMs had no practical advantages over GMMs. This finding was discussed to be rooted in the fact that the data are distributed such that the prevalent clusters are located far enough from the periodic limits and their variance is such that overlap is neglibile, more precisely that $\int_{2\pi} p(x)dx \approx 1$ for GMMs and $x \in \{\delta\theta, \delta\varphi\}$. It was also discussed that the actual distribution of the data may be a result of spatial constraints during the experimental recordings. That issue is however mitigated by the fact that the model has been shown to be robust under superimposed Gaussian noise and that further recordings in less crowded scenarios led to similar spatio-orientational distributions. Nevertheless it is expected that this mainly holds for $S^{\oplus}$ whereas the distribution of $S^{\ominus}$ will change once more data will be acquired under unconstrained conditions. Future work should clarify whether the data in $S^{\ominus}$ will be more evenly distributed along with an increasing number of observations and whether e. g. fat-tailed distributions would be a more appropriate choice for modeling $S^{\ominus}$. The fact that persons tend to avoid certain configurations under normal conditions, e. g. standing very close together and facing each other, will still show in the data so that the distribution for $S^{\ominus}$ cannot be uniform or simply assume random noise. It is furthermore expected that larger datasets will attentuate the clusters in $S^{\oplus}$.
The evaluation of the proposed model has shown relatively high performance for both GMMs and SW-GMMs. Comparison with other classifiers has shown that only SVMs were en par with the proposed model which sustains the consideration that GMMs are well-suited to reflect human interaction geometry. Accuracy, precision and recall are high for GMMs and acceptable for SW-GMMs. Performance could of course be increased by maximizing the number of components, but these were deliberately kept low to avoid overfitting and to comply with the demand for a realistic and universally applicable model. In this regard it also interesting to see that a model only based on $\delta\theta$ and a signed variant of δd which only encodes whether person j is located in front or behind person i still yields reasonable performance. Based on the analysis of the relevance of each of $\delta\theta$, $\delta\varphi$ and δd by means of differential entropy it was discussed that δd is by far the most important measure, followed by $\delta\varphi$ (sic) and $\delta\theta$. The reason why $\delta\varphi$ appears to encode more information than $\delta\theta$ is two-fold: First, $\delta\theta_{ij}$ is symmetrical to $\delta\theta_{ji}$ and therefore so is the joint distribution of $\delta\theta$ and δd, and second, values of $\delta\varphi > 2 \mod 2\pi$ are rarely observed. This is an important result for mobile SSP because $\delta\theta$ and δd are much easier to measure than $\delta\varphi$.

The previous discussion and results leave a part of the question whether the data and/or the proposed model are generalizable unsolved. Starting with the manual annotation of the data one could argue that annotations by individual labelers might lead to different results. Related work has however shown that this is not the case [149]. On the other hand, the spatial constraints and selection of participants may affect the actual distribution of the data. Contrary to further experiments which were conducted during the proceedings of this thesis, related work has reported a slight impact of spatial constraints on interpersonal distance [67], which is why future work should clarify the actual relevance and the influential extent of such constraints. The present work has however discussed potential influences by personal profile parameters such as culture, gender or age, as well as by latent variables such as group size. It should be noted that although related work agrees that likewise variables may have substantial influence on the data, most results proved to be imprecise, not based on quantitative data, and sometimes contradictory. Thus a second series of experiments was conducted in order to investigate the influence of gender, followed by a re-evaluation of the first dataset with respect to group size. Both variables are quantifiable and can be considered unambigious. Evaluation of models based on the second dataset has shown distinct distributions for male-only, female-only and male-female dyads in groups of two, three or four. δd, for example, reveals characteristic differences between genders, which are yet clearly not restricted to δd. Instead, differences also show in territorial occupancies depending on age and/or cardinality. As a matter of fact though, although specific distributions show significant differences, the size of the second dataset does not allow for further generalization such as may be found in the literature, e. g. that women tend to stand closer than men. At the bottom line, the result of the gender-related evaluation is that gender certainly has a non-negligible influence on interaction geometry, but future work should design and conduct larger experiments, possibly together with researchers from socio-psychological fields. The focus should however not only be on strict separation of gender, but instead also consider mixed configurations. According to the present results, the differences are in fact greater for mixed than for same-sex groups.

Both gender and group size were controlled parameters in the second series of experiments, whereas groups of up to nine subjects formed naturally in the initial experiment. Prior evaluations have already shown that the original model is capable of discriminating $S^{\oplus}$ and $S^{\ominus}$ under varying group size. Reevaluation of the first dataset with separate models $S^{\oplus}_{n}$ and $S^{\ominus}$ for group sizes $n \in \{2, \dots, 7, 9\}$ has shown very few misclassifications for groups of up to four persons whereas performance deteriorates quickly for larger groups. It was discussed that this is a consequence of increasing variance and changing distribution of the variables along with increasing cardinality. Smaller groups have more flexibility e. g. in terms of adapting very distinct spatio-orientational arrangements (F-formations) where each different choice itself implies overall variance in the data. Larger groups are less flexible in their choice of arrangement, but variance and overlap are generally higher e. g. due to increased distance. For example, slight changes in $\delta\theta$ may cause large variations in $\delta\varphi$ at greater distances. This reasoning is corroborated by the finding that most misclassifications occurred in favour of neighbouring classes. In order to see how else the model could profit from individual models per group size, the evaluation results of the $S^{\oplus}_{n}$ were combined into a single virtual class $S^{\oplus}_{combined}$ which led to an increase in precision (albeit at the cost of recall) when comparing the performances for $S^{\oplus}_{combined}$ with those for $S^{\oplus}$ in the original model. This matter could be further investigated by further work, especially since larger groups were not observed as often as smaller groups during the experiments. It should also be mentioned that the apparent “bias” on smaller groups was criticized by [292] with respect to the corresponding publication by Groh et al. [123]. The observed group sizes are however not a result of the experimental design or possible constraints, but instead follow the typical distribution of group sizes also known from related work which is based on quantitative data on much larger scale [79]. Nevertheless the exact modeling of a distribution for group size, e. g. in terms of a Poisson or fat-tailed distribution, can be considered unresolved and should be the subject of future work. It would also be interesting to see how interaction geometry can be used for *a posteriori* information about group size. For this, the present work has demonstrated a basic decision-theoretical approach which so far yields better than random but otherwise not acceptable results. As a proof of concept, payoff was determined as either unit distance to neighbouring classes or alternatively in the form of exponential decay. Future work might combine improved models for the class priors together with carefully chosen heuristics for the payoff. Improved solutions for this problem would be a valuable prospect as group size is a latent variable in negotiations about social situations among mobile agents. As far as the integration of profile and latent parameters into an algorithmical model for social interaction geometry is concerned, future work could for instance integrate categorical variables such as gender by means of an abstract decision tree where the path from the root to a leaf is determined by the values of the categorical variables and each leaf yields a respective model for the evaluation of ($\delta\theta$, $\delta\varphi$, δd). It is clear though that the size of the tree will grow exponentially with increasing numbers of variables and their domains, and so will the demand for additional training data. As a first step it is therefore important to determine an importance ranking between suitable variables for which not only their entropies but also their domains and potential encodings should be taken into account.

Together with the development of the proposed model this work has shown how orientation and position can be measured by mobile agents such as smartphones. In terms of orientation the main problem is relating the orientation of the mobile agent to the body of the user. In general the necessary transformation depends on precise knowledge of on-body location and orientation of the agent, although related work has e. g. projected acceleration measurements onto the horizontal plane as determined by PCA based on the notion that most acceleration (aside from gravitational force) occurs along a pedestrian's walking direction [179]. Determining on-body location and orientation [178, 142] as well as finding the correct transformation to relate agent and upper body was considered less restrictive in the present context. The proposed system is therefore based on a linear transformation based on training data which relate the phone's orientation to the body. For this, a Kinect system and smartphones were used to acquire a new dataset from several persons. The correlated data from both sources were then used to train a linear regression model. The resulting model uses the agent's measured attitude in conjunction with related temporal features for estimations of the relative heading about the yaw axis between the phone and the body. Integration of the temporal features has helped to reduce the residual error. The absolute body heading is determined by the sum of relative and absolute device heading. As the output heading is relative and the agent's heading is determined such that Gimbal lock is avoided, the system is invariant to changes in orientation. The system is still susceptible to changes in location so that a dedicated model will be required for each potential on-body location. According to related work only a limited set of discrete locations need to be considered [178, 150]. Eventually the system has been shown to perform with $\sigma \approx 9.7°$. It follows that a system which combines measurements from two agents will operate with $\sigma \approx 13.7°$. This result was used to evaluate the performance

of the GMM-based model from the original dataset after superimposition of corresponding Gaussian noise under which the interaction geometry model still performed very well with a mere $\sim 1.1\%$ loss in accuracy, a fact which contributes to the choice of GMMs as well as the understanding that the model is not overfitting the data.
For position measurements an ultrasound based system was presented. Since absolute positions are not required for interaction geometry this system sufficiently determines interpersonal distances (together with the possibility for low-quality estimates of $\delta\theta$ and $\delta\varphi$ as by-products). The system is comprised of wearable sensor boxes, each of which houses six ultrasound sensors arrayed such that sensing areas partially overlap. An external clock is used for accurate synchronization of time. The system was evaluated in a series of experiments with varying persons and group sizes against the infrared tracking system, yielding an residual error of $24.4 \pm 8.6cm$ for δd and rather large errors for $\delta\theta$ and $\delta\varphi$. It was argued that the mean of 24cm can be regarded as systematic error and thus be resolved. The GMM-based model for social interaction geometry was once again evaluated with superimposed Gaussian noise with and without the systematic error plus the standard deviation, where the model again performed well in spite of the superimosed noise. Using the ultrasound based estimates of $\delta\theta$ and/or $\delta\varphi$ leads to poor performance, where only the systematic correction of $\delta\theta$ yields acceptable results. The estimates of $\delta\theta$ and $\delta\varphi$ could however be used as backups or for the fusion with accurate measurements from more reliable sources (also with higher resolution), such as from the proposed system for orientation estimation. At the bottom line, this ultrasound based system should be seen as a proof-of-concept and as a means to verify the model for social interaction geometry when subject to real-world noisy distance measurements. Future work should consider using independent systems such as [229] although it should be ensured that a corresponding system should work in the inaudible range.

The next contribution of this thesis is the use of Subjective Logic (SL) for sensor fusion and the modeling of trust in a network of individual agents. Other than probability theory, SL assigns belief mass to sets of atomic events and thus allows for explicitly stating ignorance about parts of the state space (frame of discernment). It furthermore fosters the introduction of uncertainty to overcome known limitations of DST in cases of high conflict. It is arguable whether probability theory could be used instead. The latter would require a much more complex model plus *a priori* knowledge about the whole infrastructure of the system, a fact which does not seem reasonable in a highly heterogeneous MSN scenario. For applications of SL in MSN a new hierarchical sensor model of physical and logical sensors was presented, where e. g. higher level logical sensors may combine measurements from both local and remote logical and physical sensors. It was then shown how SL can be used to fusion sensors based on either interaction geometry and/or low-level audio features in order to output whether two agents were engaged in social interaction or not. Moreover it was discussed how SL could be used to model trust between agents. Due to the length of the experimental recordings as well as missing details about the personal background of the subjects and their relationships the actual modeling of trust was omitted and only fusioning was evaluated. Based on the fusion of the aforementioned logical sensors of interaction geometry and/or low-level audio features, several clustering methods based on the maximization of modularity were applied. The final evaluation results have shown that SL fusion of independent logical geometry and audio sensors yields significantly improved results over individual measurements when compared with the manual annotation of the dataset.

The hitherto results have shown that the new model for social interaction geometry can be generalized to some extent. It is nevertheless highly likely that there will be situations where such a model, which is based on *static* interaction geometry, will fail due to both static and dynamic components that could neither be anticipated nor integrated into the model. Examples were given such as a subway ride on a fully packed train, visiting the cinema, or attending the Vienna Opera Ball. Alternative forms of dynamic models were discussed based on frequency domain analysis, HMMs, or *Eigenzones* (PCA). It was argued that all methods will eventually suffer from modeling aspects and heuristic choices. As a result, a new dynamic model for the detection of mutual simultaneous and co-located activities was proposed. The corresponding definition requires that all participating persons perform the same *type* of activity, for which knowledge of the activity's semantics is not required. Information about co-activities can for instance be used for social network inferral as well as further insight into social relationships, for which a number of examples were discussed. For evaluation and training, a new dataset was recorded from the streams of numerous mobile phone sensors during several sessions with varying pairs of persons. Scriplets were used to outline the supposed activities during the sessions which took place in arbitrary (uncontrolled) environments. It was shown that the computation of *low-level* location-, motion- and audio-based features based on the pairwise but also the individual datastreams from the devices allows for highly accurate discrimination of the presence ($\mathsf{C}^{\oplus}$) or absence ($\mathsf{C}^{\ominus}$) of social co-activities. A decision tree classifier was used as it enables researchers to easily determine the importance of features and follow the decision process for selected samples. Parts of the tree can also be manually remodeled if necessary. Since decision trees do not *per se* support the integration of class priors, future work could investigate corresponding means such as described in [56]. So far the problem is alleviated by the fact that the present evaluation shows very high precision and recall for both classes. It was found that for a number of continuous input variables $\mathsf{C}^{\oplus}$ and $\mathsf{C}^{\ominus}$ are not linearly separable. If not treated with care this can lead to overfitting as well as very large trees. The proposed model therefore performs pruning together with a lower bound of samples per leaf. Different strategies for feature vector rates and window sizes were evaluated. According to the results a default feature vector rate of 2 Hz yields a good compromise between capturing reliable information and the ability to quickly react to changes in the performed activities. It was determined that window sizes should be chosen "inversely proportional to the average sensors' sampling rate" [19] in each group of sensors. Further analysis of the features proved that all feature groups contribute similarly to the overall decision process. Future work should nevertheless investigate the relevance of particular feature groups for certain activities or groups of activities. In case of missing features, e. g. due to the temporary loss of a physical or logical sensor in a real-world application, it was proposed to provide either individual models for different configurations of sensors, or perform a majority voting at the node of the tree at which processing had to stop due to the missing information.

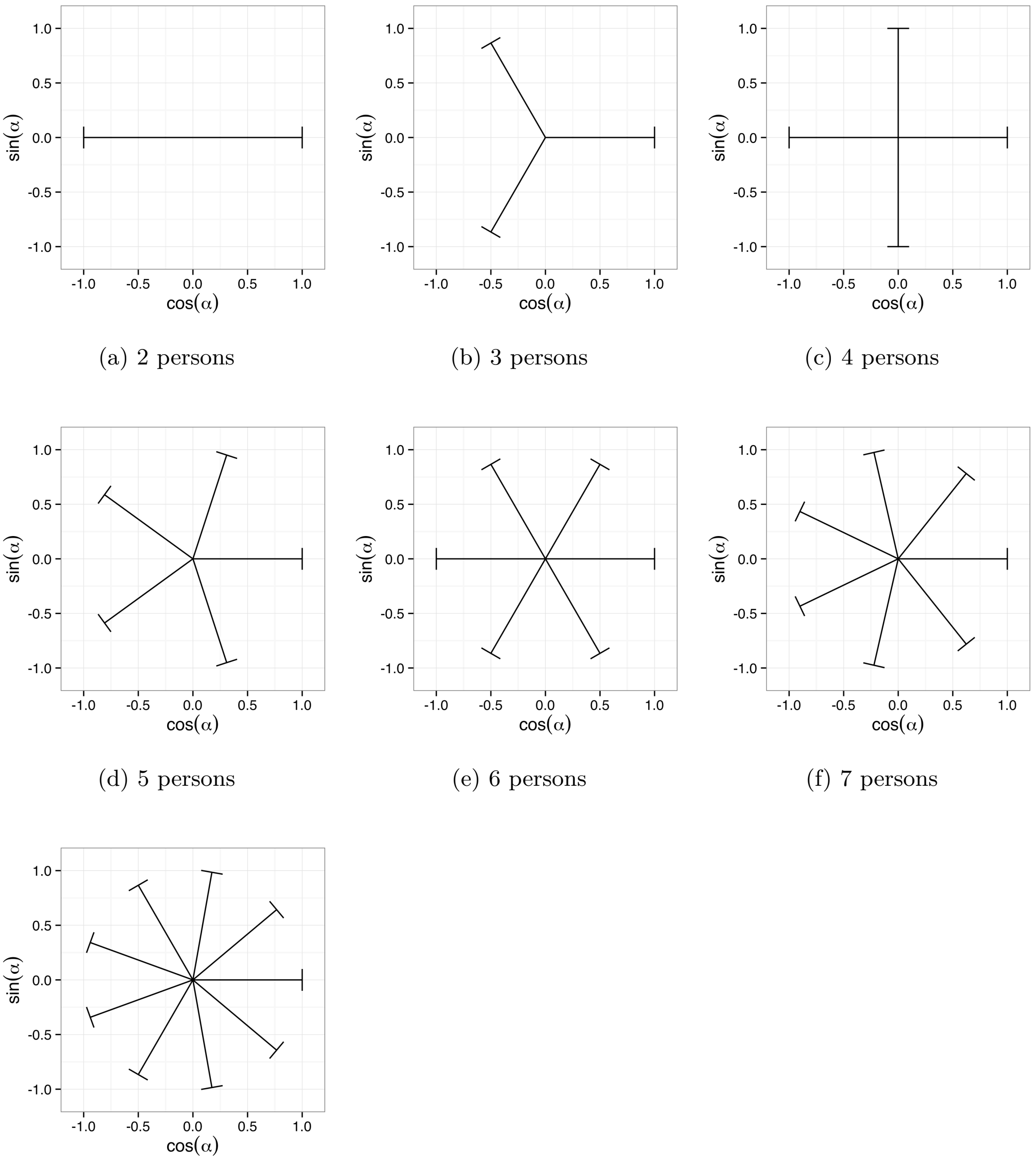
In order to determine changes in the types of co-activities a new method for the segmentation of a continuous stream of previously detected co-activities was introduced. Based on the BIC criterion which is used for a similar purpose in speaker diarization, the proposed algorithm attempts to find changing points by moving adjacent windows over the data and determining whether the data in both windows are best modeled by a single or two individual distributions. Visual inspection of the principal components of the data has shown that persons do not abruptly change their activities. Instead, observations gradually move from one cluster to another. Taking this sensitivity into account and compensating for general low detectability around frame borders, evaluation shows that the segmentation algorithm finds most true changing points but suffers from a large number of false positives. This is not unexpected since activities can be regarded at different levels in their hierarchy. The evaluation has furthermore shown that the algorithm is sensible to changes along the principal components' axes. As a number of audio-based features are close to the principal axes of the dataset, this happens for example in situations where the primary activity is suddenly accompanied by a loud noise such as a cable car passing by. It was furthermore discussed that due to the possible nesting of activities default evaluation criteria like the MDR or the FAR are not sufficient for the assessment of a co-activity segmentation algorithm. It was proposed that instead the main activity in each segment should occur in at least 90% or even 95% of the intra-segment observations which leads to significant better results for the proposed algorithm. Finding suitable measures and defining appropriate baselines for the comparison of related systems would definitely be an important point for future work.

As a consequence of the prior segmentation, the last part of the thesis was concerned with recognizing non-adjacent activities of the same type after segmentation. The proposed clustering algorithm is similar to the segmentation algorithm in that it uses the BIC criterion to decide when to join two segments whenever they are better modeled by a single than by two individual distributions. A notable difference to the former algorithm is the fact that a heuristic choice of a threshold value was necessary. While BIC usually automatically implies a corresponding threshold, that threshold had to be adapted to compensate for the smaller sample sizes after segmentation. The evaluation was performed on both the actual as well as the ideal segmentation result. Using the same performance criterion as proposed for the segmentation, the clustering algorithm shows very good performace when applied to the ideal segmentation result and acceptable performance in case of the actual segmentation result.

The latter evaluation concludes this thesis. Aside from the new contributions to the field and their careful evaluation, a number of open questions remain which were beyond the scope of this work and which were listed throughout this chapter. Readers should note that the proposed models as well as the datasets are neither claimed as exhaustive nor the final truth. Instead they are intended to serve as the basis for further research and refinement on the basis of larger scale experiments, preferably conducted by computer and social scientists alike. Most importantly, this work has shown that a significant portion of non-verbal human behaviour can be captured and recognized by universal algorithmical models such as the proposed model for social interaction geometry or the model for co-activity detection.

A

IDEAL CIRCULAR CONFIGURATIONS



(a) 2 persons

(b) 3 persons

(c) 4 persons

(d) 5 persons

(e) 6 persons

(f) 7 persons

(g) 9 persons

Figure 54.: Ideal circular configurations of varying arities.

B

SCATTER PLOTS OF THE DATA FOR $S^{\oplus}$ PER ARITY

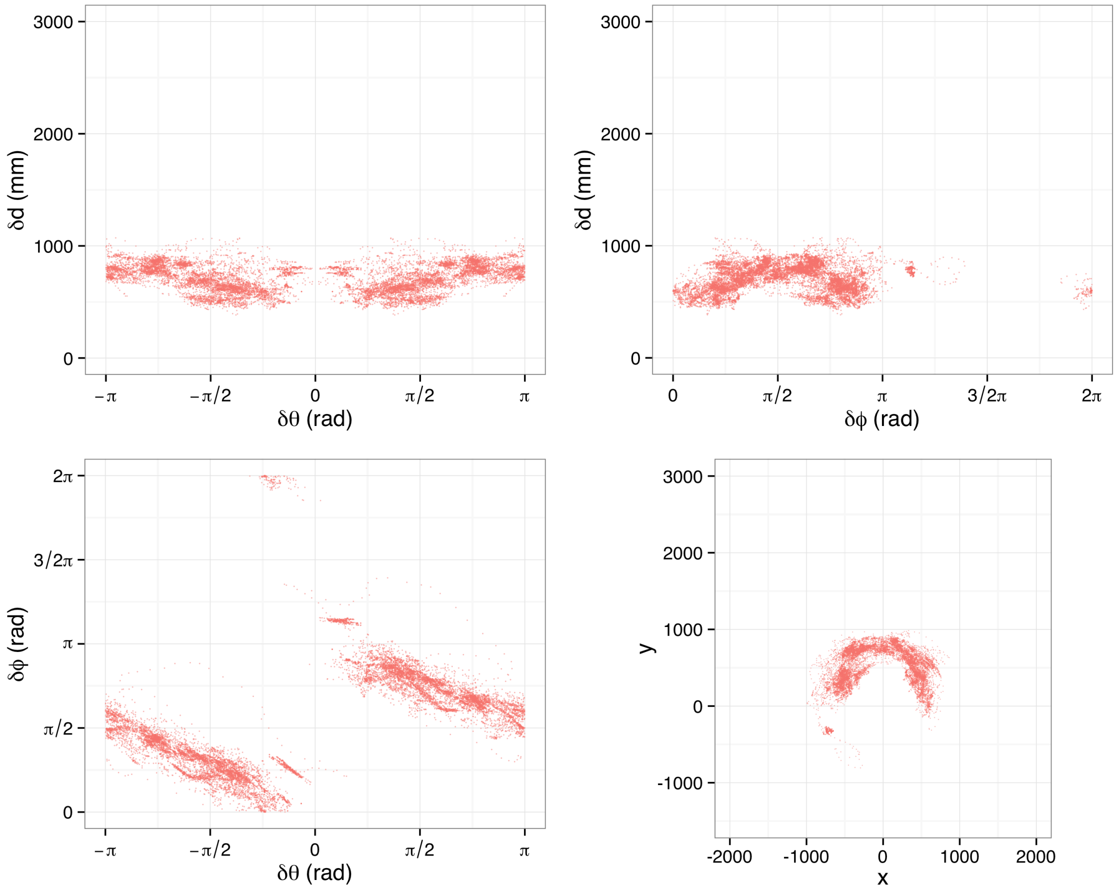


Figure 55.: Samples of $S^{\oplus}$ for social situations of two.

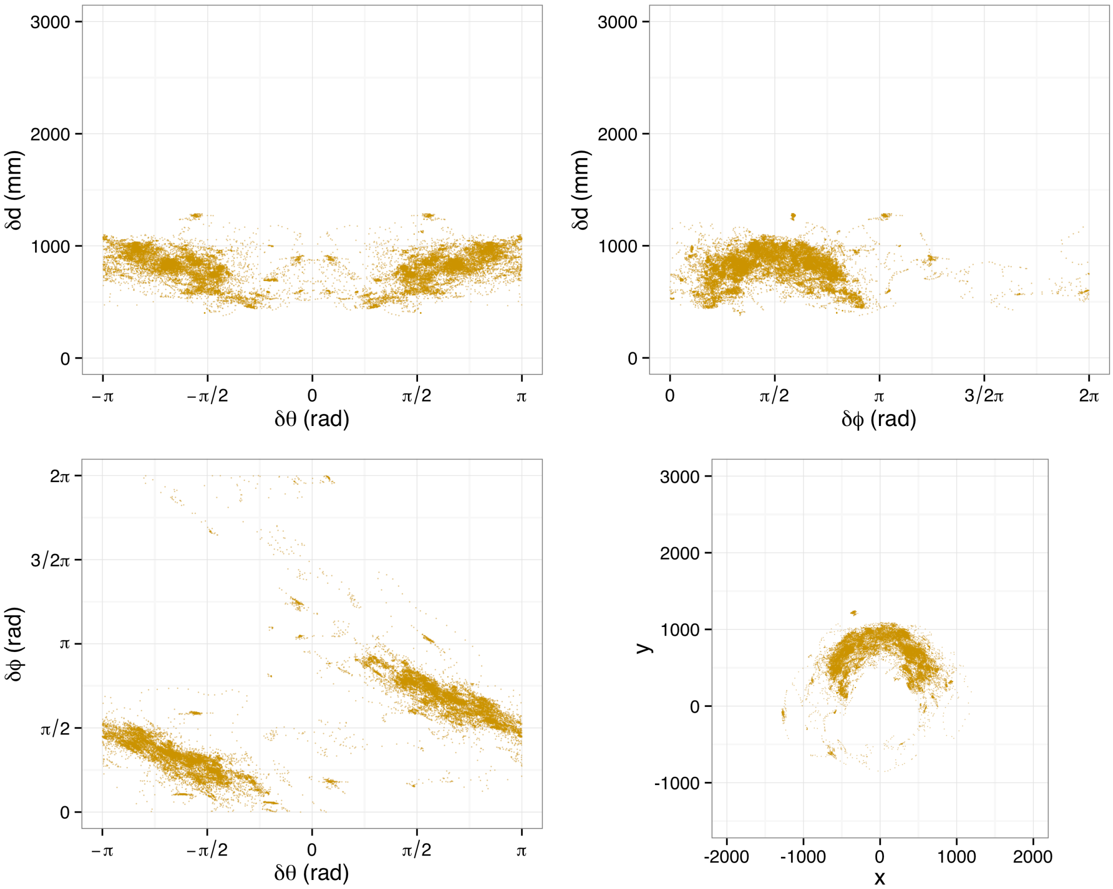


Figure 56.: Samples of $S^{\oplus}$ for social situations of three.

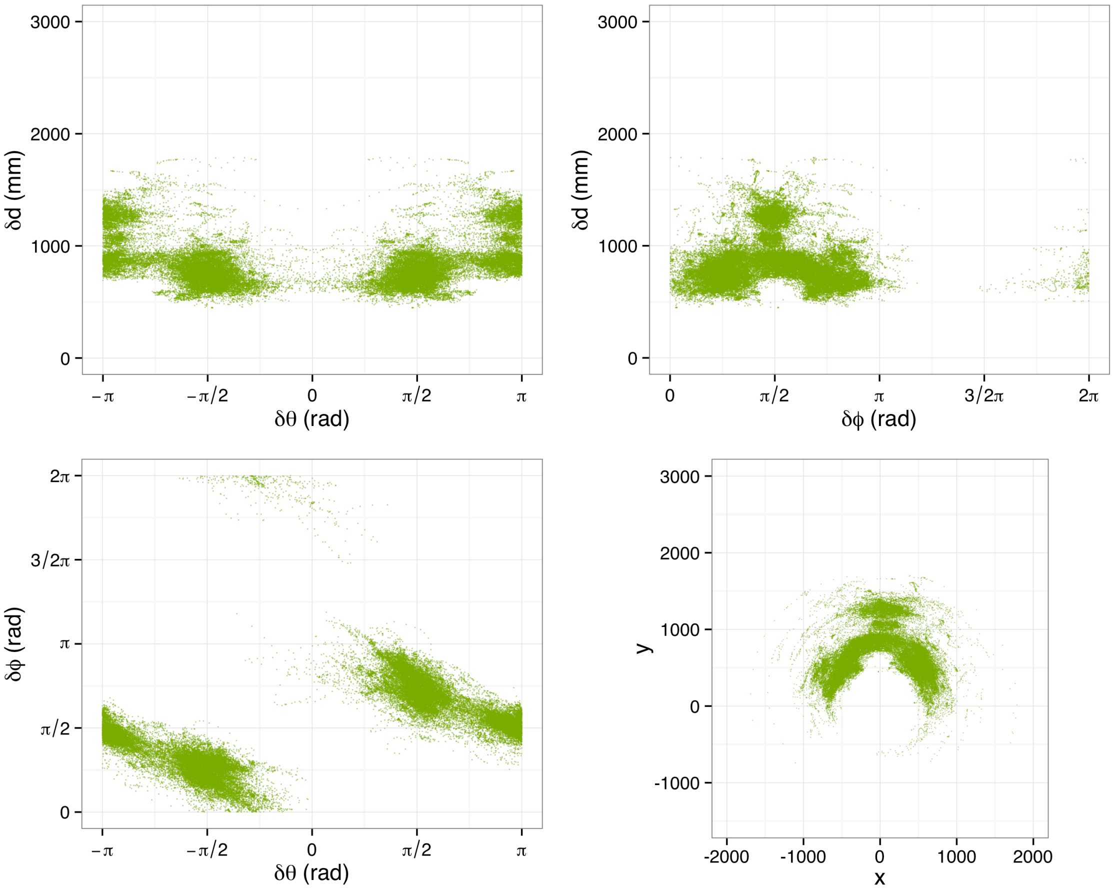


Figure 57.: Samples of $S^{\oplus}$ for social situations of four.

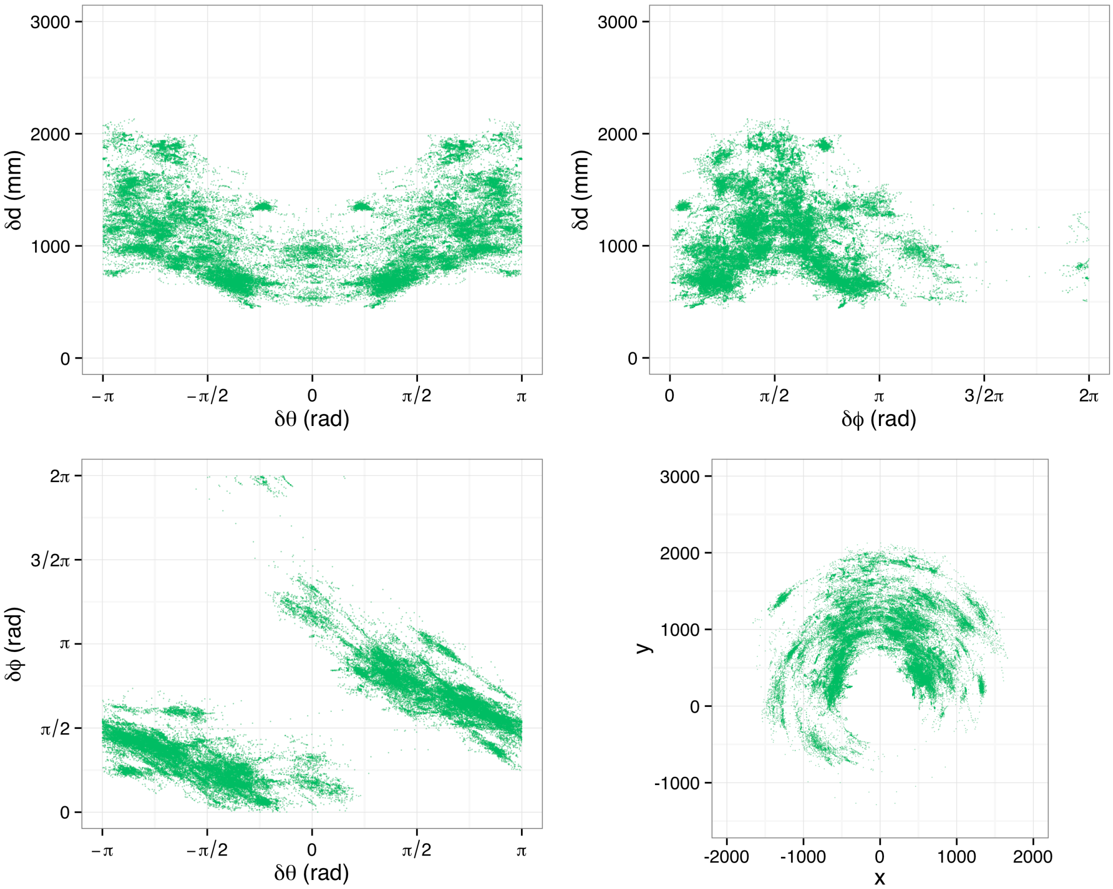


Figure 58.: Samples of $S^{\oplus}$ for social situations of five.

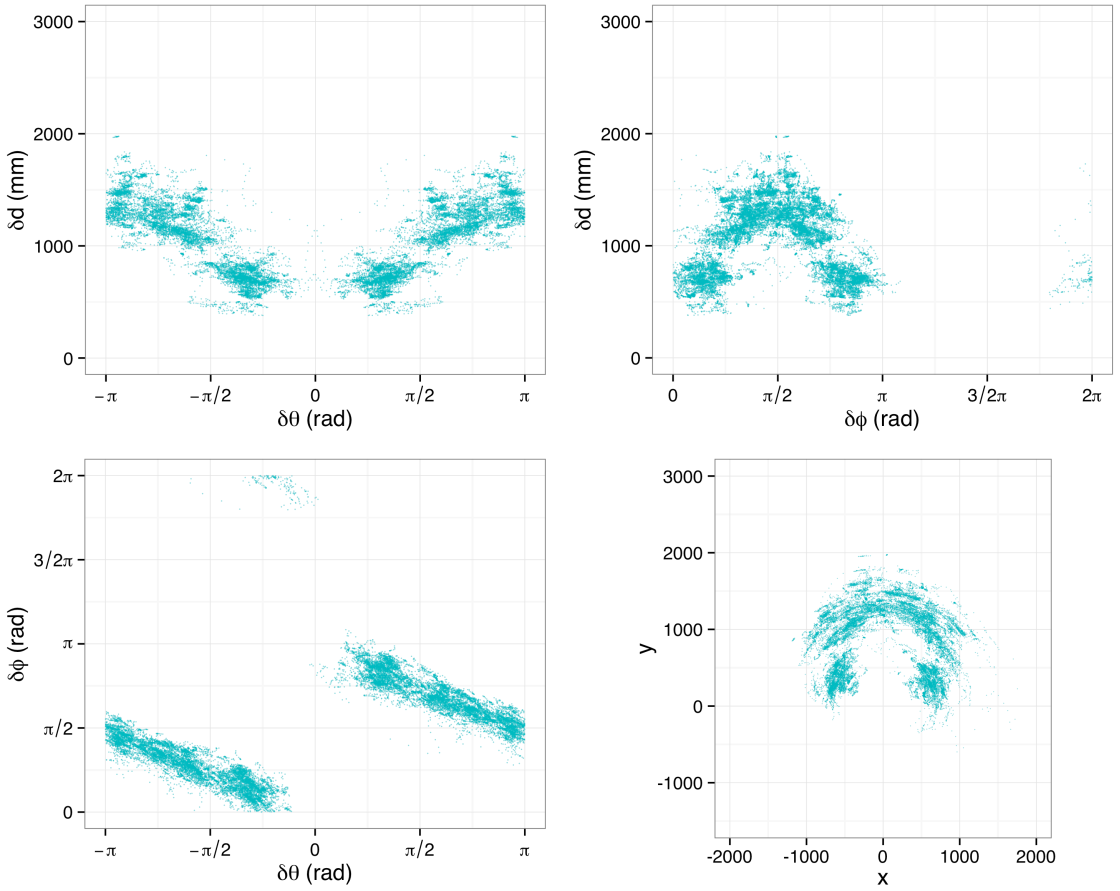


Figure 59.: Samples of $S^{\oplus}$ for social situations of six.

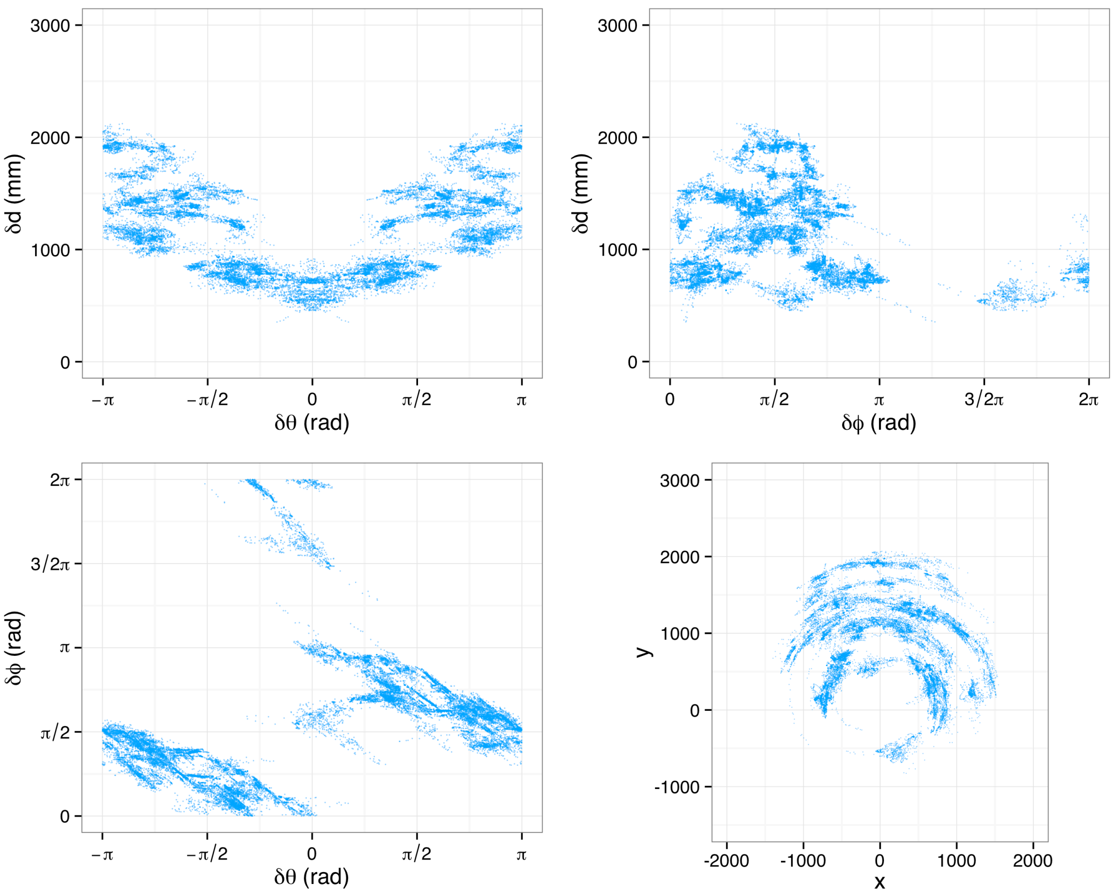


Figure 60.: Samples of $S^{\oplus}$ for social situations of seven.

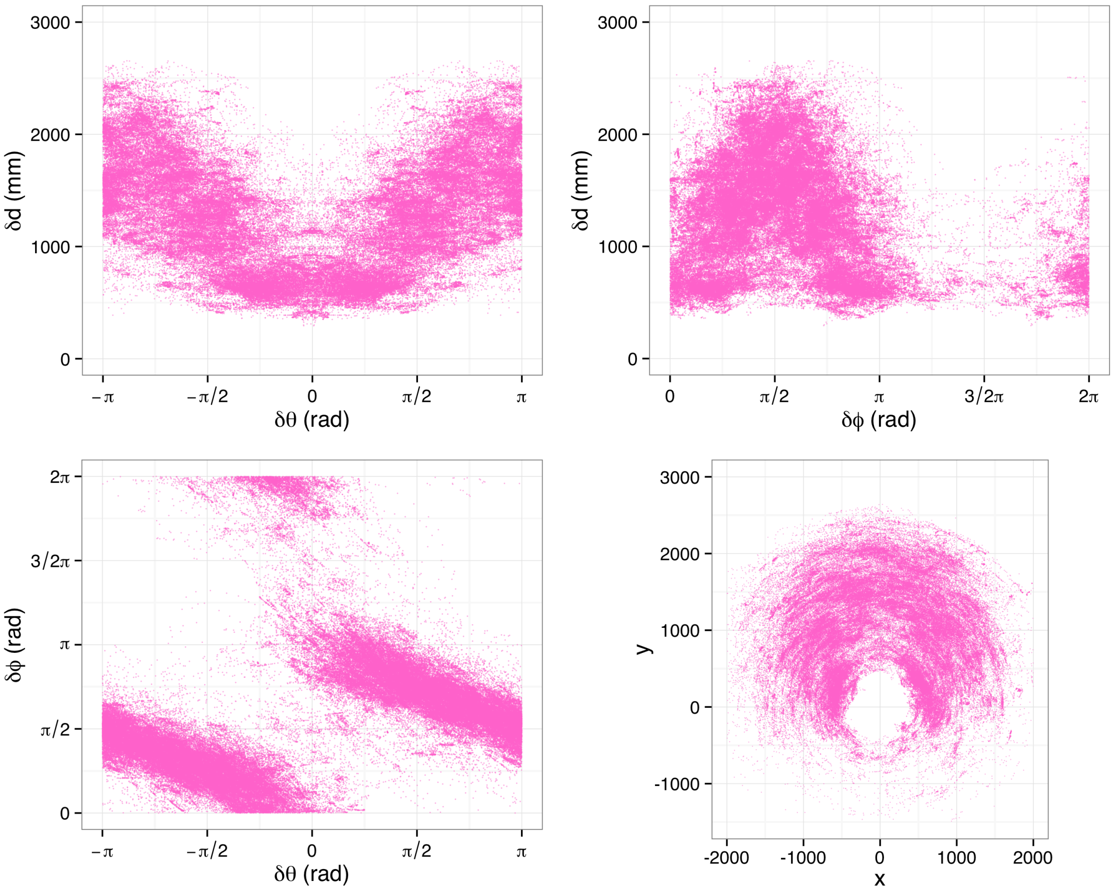


Figure 61.: Samples of $S^{\oplus}$ for social situations of nine.

C

DECISION TREE

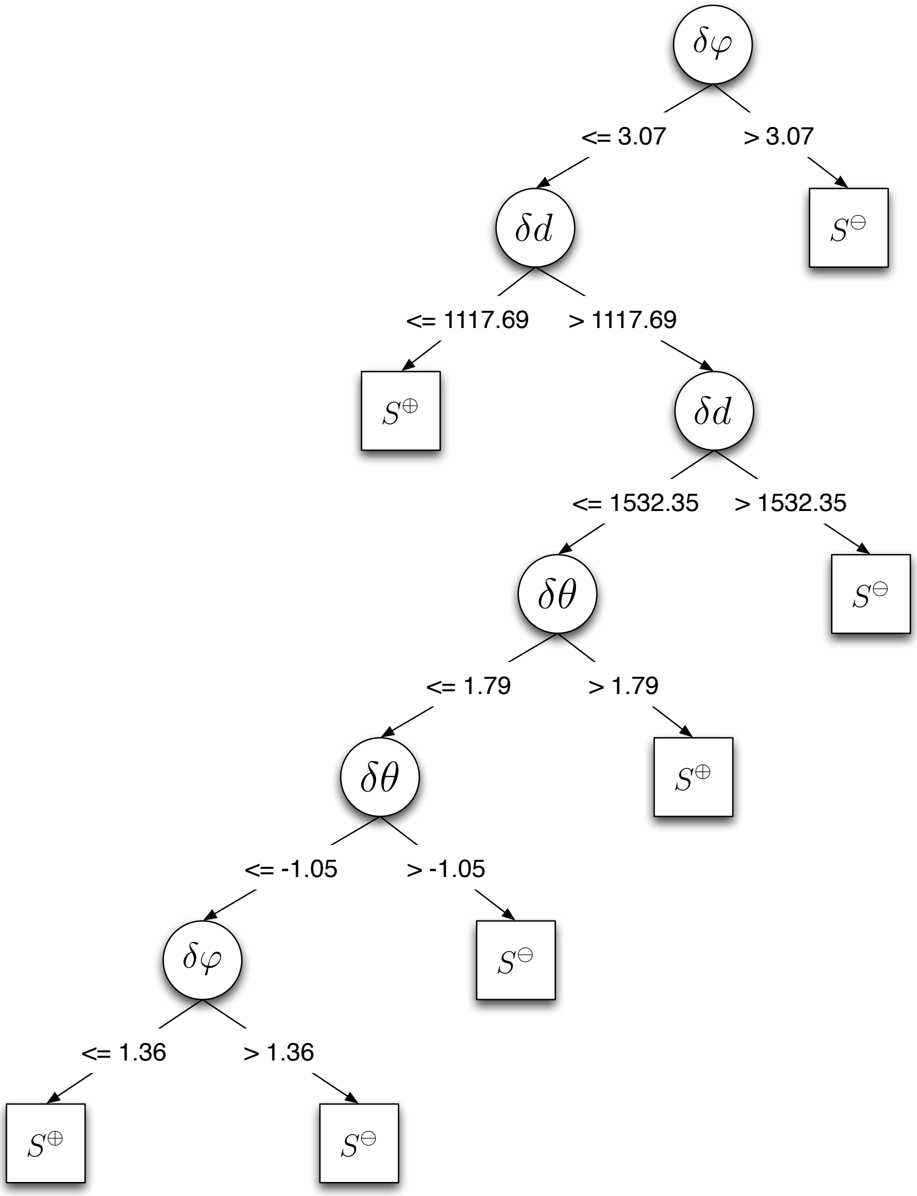


Figure 62.: Pruned decision tree with 25,000 samples per leaf as determined by J48 in [134].

D

ANNOTATION OF A RECORDED SESSION FOR CO-ACTIVITIES

```
# 2012-10-11
# Session with Daniel and Hannes around the Hohenzollernplatz.

.attributename activitytype
.defaultclass unknown
.alias daniel 63ecc0f334f552c4dcadb39959c27e2c4d462d60
.alias hannes 981d49079225398ade36eed668d26893ee1b2151

# We start from the parkbench located at Luitpoldpark.

daniel 0 10 standing
hannes 0 10 standing

daniel 10 200 walking
hannes 10 66 walking
hannes 66 302 sitting

# Hannes keeps sitting on the bench while Daniel proceeds to the tennis club
# around the corner.

# We meet again at the park bench and sit together for a short time. We then
start sportive acitivities: Throwing around fir cones! We're attacked by a dog.

daniel 200 302 sitting
daniel 302 420 walking
hannes 302 420 walking

daniel 420 580 throwingandcatching
hannes 420 580 throwingandcatching

daniel 580 9999 walking
hannes 580 9999 walking
```

Figure 63.: Example of an annotated session. Data originate from [19].

E

FEATURE WEIGHTS DURING CO-ACTIVITY DIARIZATION

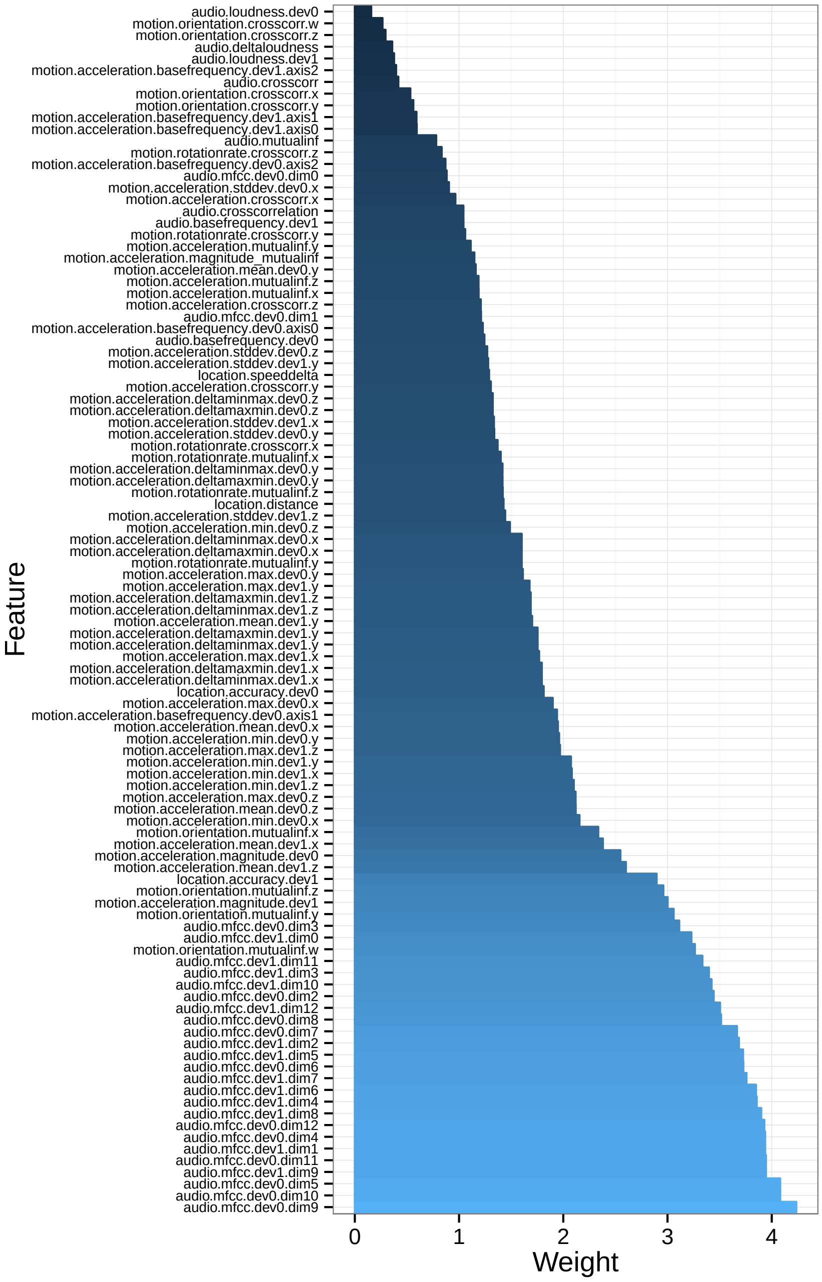


Figure 64.: Distribution of feature weights after PCA during co-activity diarization.

BIBLIOGRAPHY

[1] Kinect for Xbox 360 Hits Million Mark in Just 10 Days, November 2010. URL http://www.microsoft.com/en-us/news/press/2010/nov10/11-15ninemillionpr.aspx.

[2] Mikrozensus - Fragen zur Gesundheit - Koerpermasse der Bevoelkerung, January 2011. 5239003099004.

[3] Zensus 2011 - Bevölkerung nach Alter in Jahren und Geschlecht in Deutschland, September 2011.

[4] World Population by Age and Sex, 2013. URL http://www.census.gov.

[5] SRF02 Ultrasonic Range Finder - Technical Specification, January 2015. URL http://www.robot-electronics.co.uk/htm/srf02tech.htm.

[6] SSPNET: A European network of excellence in social signal processing, April 2015. URL http://sspnet.eu/.

[7] Mike Addlesee, Rupert Curwen, Steve Hodges, Joe Newman, Pete Steggles, Andy Ward, and Andy Hopper. Implementing a sentient computing system. *Computer*, 34(8):50–56, 2001.

[8] advanced realtime tracking GmbH. *ARTtrack & DTrack Manual v1.24.3*, July 2007.

[9] Ankur Agarwal and Bill Triggs. Recovering 3d human pose from monocular images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(1):44–58, 2006.

[10] Yannis Agiomyrgiannakis and Yannis Stylianou. Wrapped Gaussian Mixture Models for Modeling and High-Rate Quantization of Phase Data of Speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(4):775–786, May 2009. ISSN 1558-7916. doi: 10.1109/TASL.2008.2008229. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4806283.

[11] Nadav Aharony, Wei Pan, Cory Ip, Inas Khayal, and Alex Pentland. Social fMRI: Investigating and shaping social mechanisms in the real world. *Pervasive and Mobile Computing*, 7(6):643–659, December 2011. ISSN 15741192. doi: 10.1016/j.pmcj.2011.09.004. URL http://linkinghub.elsevier.com/retrieve/pii/S1574119211001246.

[12] Karl Albrecht. *Social Intelligence: The new science of success.* John Wiley & Sons, 2006.

[13] Toshitaka Amaoka, Hamid Laga, Suguru Saito, and Masayuki Nakajima. Personal space modeling for human-computer interaction. In *Entertainment Computing – ICEC 2009*, volume Lecture Notes in Computer Science Volume 5709. Springer, 2009. ISBN 978-3-642-04051-1.

[14] Ilkka Arminen and Alexandra Weilenmann. Mobile presence and intimacy—Reshaping social actions in mobile contextual configuration. *Journal of Pragmatics*, 41(10):1905–1923, 2009. URL `http://www.sciencedirect.com/science/article/pii/S0378216608002269`.

[15] Martin Atzmueller and Katy Hilgenberg. Towards capturing social interactions with sdcf: An extensible framework for mobile sensing and ubiquitous data collection. In *Proceedings of the 4th International Workshop on Modeling Social Media*, page 6. ACM, 2013.

[16] Akin Avci, Stephan Bosch, Mihai Marin-Perianu, Raluca Marin-Perianu, and Paul Havinga. Activity recognition using inertial sensing for healthcare, wellbeing and sports applications: A survey. In *Architecture of computing systems (ARCS), 2010 23rd international conference on*, pages 1–10. VDE, 2010. URL `http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5759000`.

[17] Shahid Ayub, Alireza Bahraminisaab, and Bahram Honary. A Sensor Fusion Method for Smart phone Orientation Estimation. In *Proceedings of the 13th Annual Post Graduate Symposium on the Convergence of Telecommunications, Networking and Broadcasting, Liverpool*, 2012. URL `http://www.cms.livjm.ac.uk/pgnet2012/Proceedings/Papers/1569603133.pdf`.

[18] M. Azizyan, I. Constandache, and R. Roy Choudhury. SurroundSense: mobile phone localization via ambience fingerprinting. In *Proceedings of the 15th annual international conference on Mobile computing and networking*, pages 261–272, 2009. URL `http://dl.acm.org/citation.cfm?id=1614350`.

[19] Daniel Bader. Co-activity detection with mobile sensors. Master's thesis, Technische Universität München; Supervisor: Lehmann, A. and Groh, G., 2012.

[20] Paramvir Bahl and Venkata N Padmanabhan. RADAR: An in-building RF-based user location and tracking system. In *INFOCOM 2000. Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*, volume 2, pages 775–784. IEEE, 2000.

[21] Claus Bahlmann. Directional features in online handwriting recognition. *Pattern Recognition*, 39(1):115–125, January 2006. ISSN 00313203. doi: 10.1016/j.patcog.2005.05.012. URL `http://linkinghub.elsevier.com/retrieve/pii/S0031320305002256`.

[22] Mark Baldassare and Susan Feller. Cultural variations in personal space. *Ethos*, 3 (4):481–503, 1975.

[23] Walter Bamberger, Josef Schlittenlacher, and Klaus Diepold. A trust model for inter-vehicular communication based on belief theory. In *Social Computing (SocialCom), 2010 IEEE Second International Conference on*, pages 73–80. IEEE, 2010.

[24] Udana Bandara, Mikio Hasegawa, Masugi Inoue, Hiroyuki Morikawa, and Tomonori Aoyama. Design and implementation of a bluetooth signal strength based location sensing system. In *Radio and Wireless Conference, 2004 IEEE*, pages 319–322. IEEE, 2004. URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1389140.

[25] M. Cecilia C. Baranauskas. Socially aware computing. In *ICECE 2009 VI International Conference on Engineering and Computer Education*, pages 1–5, 2009. URL http://www.researchgate.net/profile/Roberto_Pereira2/publication/234046577_SOCIALLY_AWARE_COMPUTING/links/02bfe50e86d475c653000000.pdf.

[26] Claude Barras, Xuan Zhu, Sylvain Meignier, and Jean-Luc Gauvain. Improving speaker diarization. In *RT-04F workshop*, 2004. URL ftp://tlp.limsi.fr/public/rt04f_diarization.pdf.

[27] Christopher Barthold, Kalyan Pathapati Subbu, and Ram Dantu. Evaluation of gyroscope-embedded mobile phones. In *Systems, Man, and Cybernetics (SMC), 2011 IEEE International Conference on*, pages 1632–1638. IEEE, 2011. URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6083905.

[28] James C Baxter. Interpersonal spacing in natural settings. *Sociometry*, 1970.

[29] Paul A Bell, Linda Mannik Kline, and William A Barnard. Friendship and freedom of movement as moderators of sex differences in interpersonal distancing. *The Journal of social psychology*, 128(3):305–310, 1988.

[30] James O Berger. *Statistical decision theory and Bayesian analysis*. Springer, 1985.

[31] Phyllis W Berman and Vicki L Smith. Gender and situational differences in children's smiles, touch, and proxemics. *Sex roles*, 10(5-6):347–356, 1984.

[32] J. A. Bilmes. A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. *International Computer Science Institute*, 4(510):126, 1998. URL http://lasa.epfl.ch/teaching/lectures/ML_Phd/Notes/GP-GMM.pdf.

[33] C. M. Bishop, M. Svensén, and C. K. I. Williams. GTM: The generative topographic mapping. *Neural computation*, 10(1):215–234, 1998. URL http://www.mitpressjournals.org/doi/abs/10.1162/089976698300017953.

[34] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006. ISBN 978-0387-31073-2.

[35] Christopher M Bishop, Geoffrey E Hinton, and Iain GD Strachan. GTM through time. 1997.

[36] Harold D Black. A passive system for determining the attitude of a satellite. *AIAA Journal*, 2(7):1350–1351, 1964.

[37] Ulf Blanke and Bernt Schiele. Sensing location in the pocket. *Ubicomp Poster Session*, page 2, 2008. URL `http://ulfblanke.com/research/ubicomp08/paper.pdf`.

[38] Ulf Blanke and Bernt Schiele. Towards human motion capturing using gyroscopeless orientation estimation. In *ISWC*, pages 1–2, 2010. URL `http://www.researchgate.net/publication/224204861_Towards_human_motion_capturing_using_gyroscopeless_orientation_estimation/file/72e7e51f976d70af58.pdf`.

[39] K. Boakye, B. Trueba-Hornero, O. Vinyals, and G. Friedland. Overlapped speech detection for improved speaker diarization in multiparty meetings. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 4353–4356, 2008. URL `http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4518619`.

[40] Béla Bollobás and Paul Erdős. Cliques in random graphs. *Math. Proc. Cambridge Philos. Soc*, 80(3):419–427, 1976.

[41] Gaetano Borriello, Alan Liu, Tony Offer, Christopher Palistrant, and Richard Sharp. WALRUS: wireless acoustic location with room-level resolution using ultrasound. In *Proceedings of the 3rd international conference on Mobile systems, applications, and services*, pages 191–203. ACM, 2005.

[42] Remco R Bouckaert. Naive Bayes classifiers that perform well with continuous variables. In *AI 2004: Advances in Artificial Intelligence*, pages 1089–1094. Springer, 2005.

[43] Kenneth R. Britting. *Inertial navigation system analysis.* Wiley Interscience, 1971.

[44] Andrzej K Brodzik and Robert H Enders. A case of combination of evidence in the Dempster-Shafer theory inconsistent with evaluation of probabilities. *arXiv preprint arXiv:1107.0082*, 2011.

[45] Richard R Brooks and Sundararaja S Iyengar. *Multi-sensor fusion: fundamentals and applications with software.* Prentice-Hall, Inc., 1998.

[46] Raffaele Bruno and Franca Delmastro. Design and analysis of a bluetooth-based indoor localization system. In *Personal wireless communications*, pages 711–725. Springer, 2003.

[47] Giuseppe Cardone, Andrea Cirri, Antonio Corradi, Luca Foschini, and Dario Maio. MSF: An Efficient Mobile Phone Sensing Framework. *International Journal of Distributed Sensor Networks*, 2013:1–9, 2013. ISSN 1550-1329, 1550-1477. doi: 10.1155/2013/538937. URL `http://www.hindawi.com/journals/ijdsn/2013/538937/`.

[48] GA Cavagna and P Franzetti. The determinants of the step frequency in walking in humans. *The Journal of physiology*, 373(1):235–242, 1986.

[49] A. Cavanaugh, M. Lowe, D. Cyganski, and R. J. Duckworth. WPI precision personnel locator: Inverse synthetic array reconciliation tomography. In *Position Location and Navigation Symposium (PLANS), 2012 IEEE/ION*, pages 1189–1194. IEEE, 2012. URL `http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6236974`.

[50] Eddie C. L. Chan, George Baciu, and S.C. Mak. Using Wi-Fi Signal Strength to Localize in Wireless Sensor Networks. pages 538–542. IEEE, 2009. ISBN 978-0-7695-3501-2. doi: 10.1109/CMC.2009.233. URL `http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4797055`.

[51] Eliot Dismore Chapple and Conrad Maynadier Arensberg. Measuring human relations: an introduction to the study of the interaction of individuals. *Genetic Psychology Monographs*, 1940.

[52] Sudarshan S. Chawathe. Low-latency indoor localization using bluetooth beacons. In *Intelligent Transportation Systems, 2009. ITSC'09. 12th International IEEE Conference on*, pages 1–7. IEEE, 2009. URL `http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5309711`.

[53] Scott Chen and Ponani Gopalakrishnan. Speaker, environment and channel change detection and clustering via the Bayesian Information Criterion. In *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, page 8, 1998. URL `http://www.aquaphoenix.com/presentation/candidacy/chen98.pdf`.

[54] Yongguang Chen and Hisashi Kobayashi. Signal strength based indoor geolocation. In *Communications, 2002. ICC 2002. IEEE International Conference on*, volume 1, pages 436–439. IEEE, 2002.

[55] Jaewoo Chung, Matt Donahoe, Chris Schmandt, Ig-Jae Kim, Pedram Razavai, and Micaela Wiseman. Indoor location sensing using geo-magnetism. In *Proceedings of the 9th international conference on Mobile systems, applications, and services*, pages 141–154. ACM, 2011. URL `http://dl.acm.org/citation.cfm?id=2000010`.

[56] David A Cieslak and Nitesh V Chawla. Learning decision trees for unbalanced data. In *Machine learning and knowledge discovery in databases*, pages 241–256. Springer, 2008.

[57] Joel E Cohen. *Casual groups of monkeys and men: stochastic models of elemental social systems*, volume 42. Harvard University Press, 1971.

[58] Shane Colton and F. R. C. Mentor. The balance filter, 2007. URL `http://www.filedump.net/dumped/filter1285099462.pdf`.

[59] I. Constandache, X. Bao, M. Azizyan, and R. R. Choudhury. Did you see Bob?: human localization using mobile phones. In *Proceedings of the sixteenth annual international conference on Mobile computing and networking*, pages 149–160, 2010. URL `http://dl.acm.org/citation.cfm?id=1860013`.

[60] Ionut Constandache, Shravan Gaonkar, Matt Sayler, Romit Roy Choudhury, and Landon Cox. Enloc: Energy-efficient localization for mobile phones. In *INFOCOM 2009, IEEE*, pages 2716–2720. IEEE, 2009. URL `http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5062218`.

[61] Mark Cook. Experiments on orientation and proxemics. *Human Relations*, 1970.

[62] R Dennis Cook and Sanford Weisberg. Residuals and influence in regression. 1982.

[63] Microsoft Corporation. Microsoft kinect for windows dev center, 2012. URL `http://www.microsoft.com/en-us/kinectforwindowsdev/`.

[64] Microsoft Corporation. Microsoft windows phone dev center, 2012. URL `http://dev.windowsphone.com/en-us/`.

[65] Marco Cristani. Social Computer Vision for Group Behavior Analysis. *Measuring Behavior 2012*, page 512, 2012.

[66] Marco Cristani, Loris Bazzani, Giulia Paggetti, Andrea Fossati, Diego Tosato, Alessio Del Bue, Gloria Menegaz, and Vittorio Murino. Social interaction discovery by statistical analysis of F-formations. In *BMVC*, pages 1–12, 2011. URL `https://www.vision.ee.ethz.ch/publications/papers/proceedings/eth_biwi_00863.pdf`.

[67] Marco Cristani, Giulia Paggetti, Alessandro Vinciarelli, Loris Bazzani, Gloria Menegaz, and Vittorio Murino. Towards computational proxemics: Inferring social relations from interpersonal distances. In *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*, pages 290–297. IEEE, 2011.

[68] Erik B. Dam, Martin Koch, and Martin Lillholm. Quaternions, Interpolation and Animation. Technical Report DIKU-TR-98/5, Department of Computer Science, University of Copenhagen, July 1998.

[69] Philip Daubmeier. Determining body-orientation from sensors in mobile devices. Master's thesis, Technische Universität München; Supervisor: Groh, G. and Lehmann, A., 2012.

[70] Yves-Alexandre de Montjoye, César A. Hidalgo, Michel Verleysen, and Vincent D. Blondel. Unique in the Crowd: The privacy bounds of human mobility. *Scientific Reports*, 3, March 2013. ISSN 2045-2322. doi: 10.1038/srep01376. URL `http://www.nature.com/doifinder/10.1038/srep01376`.

[71] Olivier Delalleau, Aaron Courville, and Yoshua Bengio. Efficient EM Training of Gaussian Mixtures with Missing Data. *arXiv preprint arXiv:1209.0521*, 2012. URL `http://arxiv.org/abs/1209.0521`.

[72] Arthur P Dempster. Upper and lower probabilities induced by a multivalued mapping. *The annals of mathematical statistics*, pages 325–339, 1967.

[73] Arthur P Dempster. A generalization of Bayesian inference. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 205–247, 1968.

[74] Jean Dezert, Pei Wang, and Albena Tchamova. On the validity of Dempster-Shafer theory. In *Information Fusion (FUSION), 2012 15th International Conference on*, pages 655–660. IEEE, 2012. URL `http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6289865`.

[75] Hamdi Dibeklioglu, Roberto Valenti, Albert Ali Salah, and Theo Gevers. Eyes do not lie: spontaneous versus posed smiles. In *Proceedings of the international conference on Multimedia*, pages 703–706. ACM, 2010. URL `http://dl.acm.org/citation.cfm?id=1874056`.

[76] Thomas G. Dietterich. Machine learning for sequential data: A review. In *Structural, syntactic, and statistical pattern recognition*, pages 15–30. Springer, 2002. URL `http://link.springer.com/chapter/10.1007/3-540-70659-3_2`.

[77] John M. Digman. Personality structure: Emergence of the five-factor model. *Annual review of psychology*, 41(1):417–440, 1990. URL `http://www.annualreviews.org/doi/pdf/10.1146/annurev.ps.41.020190.002221`.

[78] Michael A Dosey and Murray Meisels. Personal space and self-protection. *Journal of Personality and Social Psychology*, 11(2):93, 1969.

[79] R. I. M. Dunbar, N. D. C. Duncan, and Daniel Nettle. Size and structure of freely forming conversational groups. *Human nature*, 6(1):67–78, 1995. URL `http://link.springer.com/article/10.1007/BF02734136`.

[80] Robin IM Dunbar. Coevolution of neocortical size, group size and language in humans. *Behavioral and brain sciences*, 16(4):681–693, 1993.

[81] Dora Dzvonyar. Influence of profile parameters on social situation models, October 2013.

[82] Nathan Eagle and Alex Pentland. Reality mining: sensing complex social systems. *Personal and ubiquitous computing*, 10(4):255–268, 2006.

[83] Nathan Eagle and Alex Sandy Pentland. Eigenbehaviors: identifying structure in routine. *Behavioral Ecology and Sociobiology*, 63(7):1057–1066, April 2009. ISSN 0340-5443, 1432-0762. doi: 10.1007/s00265-009-0739-0. URL `http://www.springerlink.com/index/10.1007/s00265-009-0739-0`.

[84] Nathan Norfleet Eagle. *Machine perception and learning of complex social systems*. PhD thesis, Citeseer, 2005. URL `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.90.9810&rep=rep1&type=pdf`.

[85] Julian J Edney and Nancy L Jordan-Edney. Territorial spacing on a beach. *Sociometry*, 1974.

[86] Paul Ekman and Erika L Rosenberg. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, 1997.

[87] Paul Erdős and Alfréd Rényi. On random graphs I. *Publ. Math. Debrecen*, 6:290–297, 1959.

[88] Paul Erdős and Alfréd Rényi. On the evolution of random graphs. *Magyar Tud. Akad. Mat. Kutató Int. Közl*, 5:17–61, 1960.

[89] Frederick Erickson and Jeffrey Schultz. When is a context? Some issues and methods in the analysis of social competence. *Mind, culture, and activity: Seminal papers from the laboratory of comparative human cognition*, pages 22–31, 1997.

[90] Gary W Evans and Roger B Howard. Personal space. *Psychological bulletin*, 80(4):334, 1973.

[91] Gary W Evans, Stephen J Lepore, and Karen Mata Allen. Cross-cultural differences in tolerance for crowding: fact or fiction? *Journal of Personality and Social Psychology*, 79(2):204, 2000.

[92] Frédéric Evennou and François Marx. Advanced Integration of WiFi and Inertial Navigation Systems for Indoor Mobile Positioning. *EURASIP Journal on Advances in Signal Processing*, 2006:1–12, 2006. ISSN 1687-6172, 1687-6180. doi: 10.1155/ASP/2006/86706. URL `http://asp.eurasipjournals.com/content/2006/1/086706`.

[93] Jay A. Farrell. Computation of the Quaternion from a Rotation Matrix, 2008. URL `https://www.ee.ucr.edu/~farrell/AidedNavigation/D_App_Quaternions/Rot2Quat.pdf`.

[94] Hany Ferdinando, Handry Khoswanto, and Djoko Purwanto. Embedded Kalman filter for Inertial Measurement Unit (IMU) on the Atmega8535. In *Innovations in Intelligent Systems and Applications (INISTA), 2012 International Symposium on*, pages 1–5. IEEE, 2012. URL `http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6246978`.

[95] Viacheslav Filonenko, Charlie Cullen, and James Carswell. Investigating ultrasonic positioning on mobile phones. In *Indoor Positioning and Indoor Navigation (IPIN), 2010 International Conference on*, pages 1–8. IEEE, 2010.

[96] Nicholas I Fisher. *Statistical analysis of circular data.* Cambridge University Press, 1995.

[97] Joseph P Forgas. The perception of social episodes: Categorical and dimensional representations in two different social milieus. *Journal of Personality and Social Psychology*, 34(2):199, 1976.

[98] Eric Foxlin. Pedestrian tracking with shoe-mounted inertial sensors. *Computer Graphics and Applications, IEEE*, 25(6):38–46, 2005. URL `http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1528431`.

[99] Mark G Frank and Paul Ekman. Appearing truthful generalizes across different deception situations. *Journal of personality and social psychology*, 86(3):486, 2004.

[100] Christoph Fuchs. Kombination multipler evidenzen zur detektierung sozialer situationen. Master's thesis, Technische Universität München; Supervisor: Groh, G., 2010.

[101] K. Fukunaga and L. Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on Information Theory*, 21(1):32–40, January 1975. ISSN 0018-9448. doi: 10.1109/TIT.1975.1055330.

[102] Marco Gaertler. Clustering. In *Network analysis*, pages 178–215. Springer, 2005.

[103] Mark Gales and Steve Young. The application of hidden Markov models in speech recognition. *Foundations and Trends in Signal Processing*, 1(3):195–304, 2007.

[104] Carlos E. Galván-Tejada, José C. Carrasco-Jimenez, and Ramon Brena. Location Identification Using a Magnetic-field-based FFT Signature. *Procedia Computer Science*, 19:533–539, 2013. URL `http://www.sciencedirect.com/science/article/pii/S1877050913006790`.

[105] Raghu K. Ganti, Fan Ye, and Hui Lei. Mobile crowdsensing: current state and future challenges. *Communications Magazine, IEEE*, 49(11):32–39, 2011. URL `http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6069707`.

[106] Shravan Gaonkar, Jack Li, Romit Roy Choudhury, Landon Cox, and Al Schmidt. Micro-blog: sharing and querying content through mobile phones and social participation. In *Proceedings of the 6th international conference on Mobile systems, applications, and services*, pages 174–186, 2008. URL `http://dl.acm.org/citation.cfm?id=1378620`.

[107] Daniel Gatica-Perez. Automatic nonverbal analysis of social interaction in small groups: A review. *Image and Vision Computing*, 27(12):1775–1787, 2009. ISSN 02628856. doi: 10.1016/j.imavis.2009.01.004. URL `http://linkinghub.elsevier.com/retrieve/pii/S0262885609000109`.

[108] Dariu M Gavrila and Larry S Davis. 3-D model-based tracking of humans in action: A multi-view approach. In *Computer Vision and Pattern Recognition, 1996. Proceedings CVPR'96, 1996 IEEE Computer Society Conference on*, pages 73–80. IEEE, 1996.

[109] Hans W Gellersen, Albercht Schmidt, and Michael Beigl. Multi-sensor context-awareness in mobile devices and smart artifacts. *Mobile Networks and Applications*, 7(5):341–351, 2002.

[110] Robert Gifford. *Environmental Psychology: Principles and Practices.* Allyn & Bacon, 2nd ed. edition, 1997. ISBN 978-0205189410.

[111] Robert Gifford and Brian O'Connor. Nonverbal intimacy: clarifying the role of seating distance and orientation. *Journal of nonverbal behavior*, 10(4):207–214, 1986.

[112] Jim Giles. Inside the race to hack the Kinect. *New Scientist*, 208(2789):22–23, 2010.

[113] R. Girshick, J. Shotton, P. Kohli, Antonio Criminisi, and A. Fitzgibbon. Efficient regression of general-activity human poses from depth images. In *2011 IEEE International Conference on Computer Vision (ICCV)*, pages 415–422, November 2011. doi: 10.1109/ICCV.2011.6126270.

[114] Erving Goffman. *Behavior in public places.* New York: The Free Press, 1963.

[115] Gerard Goggin. *Cell phone culture: Mobile technology in everyday life.* Routledge, 2012.

[116] J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov. Neighbourhood components analysis. 2004. URL `http://eprints.pascal-network.org/archive/00001570/`.

[117] Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.

[118] Mark Granovetter. The strength of weak ties: A network theory revisited. *Sociological theory*, 1(1):201–233, 1983.

[119] Georg Groh. Groups and Group-Instantiations in Mobile Communities - Detection, Modeling and Applications. 2007.

[120] Georg Groh. *Contextual Social Networking.* PhD thesis, TU München, 2011. URL `http://d-nb.info/1031076115/34`.

[121] Georg Groh and Alexander Lehmann. A New Data-Set for Research on Audio-Detection and Modeling of Social Micro-Contexts. Technical Report TUM-I1011, Technische Universität München, 2010.

[122] Georg Groh and Alexander Lehmann. Deducing evidence for social situations from dynamic geometric interaction data. *International Journal of Social Computing and Cyber-Physical Systems*, 1(2):206–222, 2011. URL `http://inderscience.metapress.com/index/H352050203381636.pdf`.

[123] Georg Groh, Alexander Lehmann, Jonas Reimers, Marc René Frieß, and Loren Schwarz. Detecting social situations from interaction geometry. In *2010 IEEE Second International Conference on Social Computing (SocialCom)*, pages 1–8. IEEE, 2010. URL `http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5590934`.

[124] Georg Groh, Alexander Lehmann, Tianyu Wang, Stefan Huber, and Felix Hammerl. Applications for social situation models. In *International Conference on Wireless Applications and Computing Conference, Freiburg, Germany*. Citeseer, 2010. URL `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.177.5609&rep=rep1&type=pdf`.

[125] Georg Groh, Christoph Fuchs, and Alexander Lehmann. Combining evidence for social situation detection. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, pages 742–747. IEEE, 2011.

[126] Georg Groh, Alexander Lehmann, and Manuel de Souza. Mobile detection of social situations with turn-taking patterns. *M CC SSI MCCSIS*, page 137, 2011. URL `http://ims.mii.lt/ims/konferenciju_medziaga/MCCSIS/I_WAC_TNS_2011.pdf#page=158`.

[127] Karsten Großekatthöfer and Zizung Yoon. Introduction into quaternions for spacecraft attitude representation, May 2012. URL `http://www.tu-berlin.de/fileadmin/fg169/miscellaneous/Quaternions.pdf`.

[128] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182, 2003. URL `http://dl.acm.org/citation.cfm?id=944968`.

[129] Edward T Hall. The anthropology of manners. *Scientific American*, 192:84–91, 1955.

[130] Edward T Hall. *The silent language*, volume 3. Doubleday New York, 1959.

[131] Edward T Hall. A system for the notation of proxemic behavior. *American anthropologist*, 65(5):1003–1026, 1963.

[132] Edward T Hall. Proxemics. *Current anthropology*, pages 83–108, 1968.

[133] Edward T Hall. *The Hidden Dimension.* Anchor Books New York, 1969.

[134] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18, 2009.

[135] Ismail Haritaoglu. A real time system for detection and tracking of people and recognizing their activities. *PhD Proposal, University of Maryland at College Park*, 1998.

[136] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*, volume 2. Cambridge Univ Press, 2000.

[137] John J Hartnett, Kent G Bailey, and Frank W Gibson Jr. Personal space as influenced by sex and type of movement. *The Journal of psychology*, 76(2):139–144, 1970.

[138] M. Hazas, C. Kray, H. Gellersen, H. Agbota, G. Kortuem, and A. Krohn. A relative positioning system for co-located mobile devices. In *Proceedings of the 3rd international conference on Mobile systems, applications, and services*, pages 177–190, 2005. URL `http://dl.acm.org/citation.cfm?id=1067170.1067190`.

[139] RV Heckel and JM Hiers. Social distance and locus of control. *Journal of Clinical Psychology*, 1977.

[140] Heini Hediger et al. Wild animals in captivity: An outline of the biology of zoological gardens. *Wild animals in captivity. An outline of the biology of zoological gardens.*, 1950.

[141] Jon C Helton. Uncertainty and sensitivity analysis in the presence of stochastic and subjective uncertainty. *Journal of Statistical Computation and Simulation*, 57(1-4): 3–76, 1997.

[142] Apiwat Henpraserttae, Surapa Thiemjarus, and Sanparith Marukatat. Accurate Activity Recognition Using a Mobile Phone Regardless of Device Orientation and Location. pages 41–46. IEEE, May 2011. ISBN 978-1-4577-0469-7. doi: 10.1109/BSN.2011.8. URL `http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5955295`.

[143] Stanley Heshka and Yona Nelson. Interpersonal speaking distance as a function of age, sex, and relationship. *Sociometry*, 1972.

[144] Seyed Amir Hoseini-Tabatabaei, Alexander Gluhak, and Rahim Tafazolli. A survey on smartphone-based systems for opportunistic user context recognition. *ACM Computing Surveys*, 45(3):1–51, June 2013. ISSN 03600300. doi: 10.1145/2480741.2480744. URL `http://dl.acm.org/citation.cfm?doid=2480741.2480744`.

[145] Harold Hotelling. The Generalization of Student's Ratio. *The Annals of Mathematical Statistics*, 2(3):360–378, 1931. doi: 10.1214/aoms/1177732979. URL `http://dx.doi.org/10.1214/aoms/1177732979`.

[146] Polly Huang. Promoting wearable computing: A survey and future agenda. 2000. URL `http://e-collection.library.ethz.ch/view/eth:24765`.

[147] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985.

[148] Pan Hui, Eiko Yoneki, Shu Yan Chan, and Jon Crowcroft. Distributed community detection in delay tolerant networks. In *Proceedings of 2nd ACM/IEEE international workshop on Mobility in the evolving internet architecture*, page 7. ACM, 2007.

[149] Hayley Hung and Ben Kröse. Detecting F-formations as dominant sets. In *Proceedings of the 13th international conference on multimodal interfaces*, pages 231–238, 2011. URL `http://dl.acm.org/citation.cfm?id=2070525`.

[150] Fumiko Ichikawa, Jan Chipchase, and Raphael Grignani. Where's the phone? A study of mobile phone location in public spaces. 2005. URL `http://digital-library.theiet.org/content/conferences/10.1049/cp_20051557`.

[151] Nacim Ihaddadene and Chabane Djeraba. Real-time crowd motion analysis. In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pages 1–4. IEEE, 2008. URL `http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4761041`.

[152] John James. The distribution of free-forming small group size. *American Sociological Review*, 1953.

[153] Dusan Jan, David Herrera, Bilyana Martinovski, David G Novick, and David Traum. A computational model of culture-specific conversational behavior. 2007.

[154] Edwin T Jaynes. *Probability theory: the logic of science.* Cambridge university press, 2003.

[155] Yunye Jin, Hong-Song Toh, Wee-Seng Soh, and Wai-Choong Wong. A robust dead-reckoning pedestrian tracking system with low cost sensors. In *Pervasive Computing and Communications (PerCom), 2011 IEEE International Conference on*, pages 222–230. IEEE, 2011. URL `http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5767590`.

[156] R. Jirawimut, P. Ptasinski, V. Garaj, F. Cecelja, and W. Balachandran. A method for dead reckoning parameter correction in pedestrian navigation system. *IEEE Transactions on Instrumentation and Measurement*, 52(1):209–215, February 2003. ISSN 0018-9456. doi: 10.1109/TIM.2002.807986. URL `http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1191431`.

[157] Stanley E Jones. A comparative proxemics analysis of dyadic interaction in selected subcultures of New York City. *The Journal of Social Psychology*, 84(1):35–44, 1971.

[158] Thomas Judd. A personal dead reckoning module. In *ION GPS*, volume 97, pages 47–51, 1997.

[159] Audun Jøsang. A logic for uncertain probabilities. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 9(03):279–311, 2001. URL http://www.worldscientific.com/doi/pdf/10.1142/S0218488501000831.

[160] Audun Jøsang. The consensus operator for combining beliefs. *Artificial Intelligence*, 141(1):157–170, 2002.

[161] Audun Jøsang. Probabilistic logic under uncertainty. In *Proceedings of the thirteenth Australasian symposium on Theory of computing-Volume 65*, pages 101–110. Australian Computer Society, Inc., 2007. URL http://dl.acm.org/citation.cfm?id=1273707.

[162] Audun Jøsang and Simon Pope. Dempster's rule as seen by little colored balls. *Computational Intelligence*, 28(4):453–474, 2012.

[163] Audun Jøsang, Stephen Marsh, and Simon Pope. Exploring different types of trust propagation. In *Trust management*, pages 179–192. Springer, 2006. URL http://link.springer.com/chapter/10.1007/11755593_14.

[164] Abhishek Kar. Skeletal tracking using microsoft kinect. *Methodology*, 1:1–11, 2010. URL http://www.cs.berkeley.edu/~akar/cs397/Skeletal%20Tracking%20Using%20Microsoft%20Kinect.pdf.

[165] Adam Kendon. Some functions of gaze-direction in social interaction. *Acta psychologica*, 26:22–63, 1967.

[166] Adam Kendon. *Conducting interaction: Patterns of behavior in focused encounters*. Cambridge University Press, 1990.

[167] Adam Kendon. Spacing and orientation in co-present interaction. In *Development of Multimodal Interfaces: Active Listening and Synchrony*, pages 1–15. Springer, 2010.

[168] Kourosh Khoshelham. Accuracy Analysis of Kinect Depth Data. In *ISPRS workshop laser scanning*, volume 38, page W12, 2011. URL http://www.int-arch-photogramm-remote-sens-spatial-inf-sci.net/XXXVIII-5-W12/133/2011/isprsarchives-XXXVIII-5-W12-133-2011.pdf.

[169] Kourosh Khoshelham and Sander Oude Elberink. Accuracy and Resolution of Kinect Depth Data for Indoor Mapping Applications. *Sensors*, 12(12):1437–1454, February 2012. ISSN 1424-8220. doi: 10.3390/s120201437. URL http://www.mdpi.com/1424-8220/12/2/1437/.

[170] A. D. King. Inertial Navigation - Forty Years of Evolution. *GEC review*, 13(3):140–149, 1998. URL http://www.imar-navigation.de/downloads/papers/inertial_navigation_introduction.pdf.

[171] David Kleinbaum, Lawrence Kupper, Azhar Nizam, and Eli Rosenberg. *Applied regression analysis and other multivariable methods.* Cengage Learning, 2013.

[172] A. R. Klumpp. Singularity-free extraction of a quaternion from a direction-cosine matrix. *Journal of Spacecraft and Rockets*, 13(12):754–755, December 1976. ISSN 0022-4650, 1533-6794. doi: 10.2514/3.27947. URL http://arc.aiaa.org/doi/abs/10.2514/3.27947.

[173] Mark Knapp, Judith Hall, and Terrence Horgan. *Nonverbal communication in human interaction.* Cengage Learning, 2013.

[174] Nisarg Kothari. *An Extended Kalman Filter for Cell Phone Orientation Tracking.* 2011. URL http://www.andrew.cmu.edu/user/ndk/KDC_Report.pdf.

[175] M. Kotti, V. Moschou, and C. Kotropoulos. Speaker segmentation and clustering. *Signal processing*, 88(5):1091–1124, 2008. URL http://www.sciencedirect.com/science/article/pii/S016516840700391X.

[176] Jack B Kuipers. *Quaternions and rotation sequences.* Princeton university press Princeton, 1999.

[177] Ziva Kunda. *Social cognition: Making sense of people.* MIT press, 1999.

[178] Kai Kunze, Paul Lukowicz, Holger Junker, and Gerhard Tröster. Where am I: Recognizing On-Body Positions of Wearable Sensors. In *Location-and Context-Awareness*, pages 264–275. Springer, 2005. URL http://link.springer.com/chapter/10.1007/11426646_25.

[179] Kai Kunze, Paul Lukowicz, Kurt Partridge, and Bo Begole. Which Way Am I Facing: Inferring Horizontal Device Orientation from an Accelerometer Signal. pages 149–150. IEEE, September 2009. ISBN 978-0-7695-3779-5. doi: 10.1109/ISWC.2009.33. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5254664.

[180] Francesco Lagona and Marco Picone. Model-based clustering of multivariate skew data with circular components and missing values. *Journal of Applied Statistics*, 39 (5):927–945, 2012.

[181] Nicholas D. Lane, Emiliano Miluzzo, Hong Lu, Daniel Peebles, Tanzeem Choudhury, and Andrew T. Campbell. A survey of mobile phone sensing. *Communications Magazine, IEEE*, 48(9):140–150, 2010. URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5560598.

[182] Oscar D. Lara and Miguel A. Labrador. A Survey on Human Activity Recognition using Wearable Sensors. *IEEE Communications Surveys & Tutorials*, 15(3):1192–1209, 2013. ISSN 1553-877X. doi: 10.1109/SURV.2012.110112.00192. URL `http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6365160`.

[183] Neil D Lawrence, Antony IT Rowstron, Christopher M Bishop, and Michael J Taylor. Optimising synchronisation times for mobile devices. *Advances in Neural Information Processing Systems*, 2:1401–1408, 2002.

[184] Guy Lebanon. Linear Regression, 2003.

[185] Yann LeCun, Léon Bottou, Genevieve B Orr, and Klaus-Robert Müller. Efficient backprop. In *Neural networks: Tricks of the trade*, pages 9–50. Springer, 1998.

[186] Daniel D Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556–562, 2001.

[187] Matthew L. Lee and Anind K. Dey. Lifelogging memory appliance for people with episodic memory impairment. In *Proceedings of the 10th international conference on Ubiquitous computing*, pages 44–53. ACM, 2008. URL `http://dl.acm.org/citation.cfm?id=1409643`.

[188] Alexander Lehmann. *Towards Mobile Location and Orientation-Based Detection of Social Situations.* mastersthesis, Technische Universitaet Muenchen, 2009.

[189] Miriam Leibman. The effects of sex and race norms on personal space. *Environment and behavior*, 1970.

[190] Fan Li, Chunshui Zhao, Guanzhong Ding, Jian Gong, Chenxing Liu, and Feng Zhao. A reliable and accurate indoor localization method using phone inertial sensors. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pages 421–430. ACM, 2012. URL `http://dl.acm.org/citation.cfm?id=2370280`.

[191] Li-Jia Li and Li Fei-Fei. What, where and who? classifying events by scene and object recognition. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE, 2007. URL `http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4408872`.

[192] J. A. B. Link, P. Smith, N. Viol, and K. Wehrle. Footpath: Accurate map-based indoor navigation using smartphones. In *Indoor Positioning and Indoor Navigation (IPIN), 2011 International Conference on*, pages 1–8, 2011. URL `http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6071934`.

[193] Kenneth B Little. Cultural variations in social schemata. *Journal of Personality and Social Psychology*, 10(1):1, 1968.

[194] Natasha Lomas. Why Qualcomm Wants To Bring Ultrasound Transmitters To Smartphones And Tablets, February 2013. URL http://techcrunch.com/2013/02/27/qualcomm-epos-ultrasound/.

[195] Paul Lukowicz, Alois Ferscha, and others. From context awareness to socially aware computing. *IEEE Pervasive Computing*, (1):32–41, 2011. URL http://www.computer.org/csdl/mags/pc/2012/01/mpc2012010032.html.

[196] John MacCormick. How does the Kinect work?, 2011.

[197] Anil Madhavapeddy and Alastair Tse. A study of bluetooth propagation using accurate indoor location mapping. In *UbiComp 2005: Ubiquitous Computing*, pages 105–122. Springer, 2005. URL http://link.springer.com/chapter/10.1007/11551201_7.

[198] David Magnusson. An analysis of situational dimensions. *Perceptual and Motor Skills*, 32(3):851–867, 1971.

[199] Tobias P Mann. Numerically stable hidden markov model implementation. Technical report, 2006. URL http://automatica.dei.unipd.it/public/Schenato/PSC/2009_2010/materiale/gruppo%204/RiferimentiBibliografici/hmm_scaling_revised.pdf.

[200] Kanti V Mardia and Peter E Jupp. *Directional statistics*, volume 494. Wiley. com, 2009.

[201] Kantilal Varichand Mardia, John T Kent, and John M Bibby. *Multivariate analysis.* Academic press, 1979.

[202] Lukas Märdian. Determining interaction geometry with ultrasound sensors, August 2013.

[203] Herbert W Marsh, Benjamin Nagengast, and Alexandre JS Morin. Measurement invariance of big-five factors over the life span: ESEM tests of gender, age, plasticity, maturity, and la dolce vita effects. *Developmental psychology*, 49(6):1194, 2013.

[204] Paul Marshall, Yvonne Rogers, and Nadia Pantidi. Using F-formations to analyse spatial patterns of interaction in physical environments. page 445. ACM Press, 2011. ISBN 9781450305563. doi: 10.1145/1958824.1958893. URL http://portal.acm.org/citation.cfm?doid=1958824.1958893.

[205] Aleksandar Matic, Venet Osmani, Alban Maxhuni, and Oscar Mayora. Multi-modal mobile sensing of social interactions. In *Pervasive Computing Technologies for Healthcare (PervasiveHealth), 2012 6th International Conference on*, pages 105–114, 2012. URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6240369.

[206] Aleksandar Matic, Venet Osmani, and Oscar Mayora-Ibarra. Analysis of Social Interactions Through Mobile Phones. *Mobile Networks and Applications*, 17(6):808–819, 2012. ISSN 1383-469X, 1572-8153. doi: 10.1007/s11036-012-0400-4. URL `http://link.springer.com/10.1007/s11036-012-0400-4`.

[207] Uwe Maurer, Asim Smailagic, Daniel P. Siewiorek, and Michael Deisher. Activity recognition and monitoring using multiple sensors on different body positions. In *Wearable and Implantable Body Sensor Networks, 2006. BSN 2006. International Workshop on*, pages 4–pp. IEEE, 2006. URL `http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1612909`.

[208] Robert R McCrae and Oliver P John. An introduction to the five-factor model and its applications. *Journal of personality*, 60(2):175–215, 1992.

[209] Geoffrey McLachlan and David Peel. *Finite mixture models*. Wiley. com, 2004.

[210] Sylvain Meignier, Daniel Moraru, Corinne Fredouille, Jean-François Bonastre, and Laurent Besacier. Step-by-step and integrated approaches in broadcast news speaker diarization. *Computer Speech & Language*, 20(2-3):303–330, April 2006. ISSN 08852308. doi: 10.1016/j.csl.2005.08.002. URL `http://linkinghub.elsevier.com/retrieve/pii/S0885230805000471`.

[211] Emiliano Miluzzo, Nicholas D. Lane, Shane B. Eisenman, and Andrew T. Campbell. CenceMe - Injecting sensing presence into social networking applications. In *Smart Sensing and Context*, pages 1–28. Springer, 2007. URL `http://link.springer.com/chapter/10.1007/978-3-540-75696-5_1`.

[212] John Mingers. An empirical comparison of pruning methods for decision tree induction. *Machine learning*, 4(2):227–243, 1989.

[213] David Mizell. Using Gravity to Estimate Accelerometer Orientation. In *Proceedings of the 7th IEEE International Symposium on Wearable Computers*, page 252. IEEE Computer Society, 2003.

[214] Thomas B Moeslund and Erik Granum. A survey of computer vision-based human motion capture. *Computer Vision and Image Understanding*, 81(3):231–268, 2001.

[215] Gelareh Mohammadi, Alessandro Vinciarelli, and Marcello Mortillaro. The voice of personality: mapping nonverbal vocal behavior into trait attributions. In *Proceedings of the 2nd international workshop on Social signal processing*, pages 17–20, 2010. URL `http://dl.acm.org/citation.cfm?id=1878123`.

[216] Desmond Morris and G Desebrock. *Manwatching: A field guide to human behaviour*. HN Abrams New York, 1977.

[217] Mehdi Moussaïd, Niriaska Perozo, Simon Garnier, Dirk Helbing, and Guy Theraulaz. The walking behaviour of pedestrian social groups and its impact on crowd dynamics. *PloS one*, 5(4):e10047, 2010.

[218] Kevin P Murphy. *Machine learning: a probabilistic perspective.* The MIT Press, 2012.

[219] Mark EJ Newman. The structure and function of complex networks. *SIAM review*, 45(2):167–256, 2003.

[220] Mark EJ Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582, 2006.

[221] Mark EJ Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004.

[222] Mark EJ Newman, Steven H Strogatz, and Duncan J Watts. Random graphs with arbitrary degree distributions and their applications. *Physical Review E*, 64(2):026118, 2001.

[223] D Warner North. A tutorial introduction to decision theory. *Systems Science and Cybernetics, IEEE Transactions on*, 4(3):200–210, 1968.

[224] Humphry Osmond. The relationship between architect and psychiatrist. *Psychiatric architecture. Washington, DC: American Psychiatric Association*, pages 16–20, 1959.

[225] A. Pantelopoulos and N.G. Bourbakis. A Survey on Wearable Sensor-Based Systems for Health Monitoring and Prognosis. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 40(1):1–12, January 2010. ISSN 1094-6977, 1558-2442. doi: 10.1109/TSMCC.2009.2032660. URL `http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5306098`.

[226] Maja Pantic, Roderick Cowie, Francesca D'Errico, Dirk Heylen, Marc Mehu, Catherine Pelachaud, Isabella Poggi, Marc Schroeder, and Alessandro Vinciarelli. Social signal processing: The research agenda. In *Visual Analysis of Humans*, pages 511–538. Springer, 2011. URL `http://link.springer.com/chapter/10.1007/978-0-85729-997-0_26`.

[227] Miles L Patterson and Lee B Sechrest. Interpersonal distance and impression formation. *Journal of Personality*, 38(2):161–166, 1970.

[228] Eric Paulos and Elizabeth Goodman. The familiar stranger: anxiety, comfort, and play in public places. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 223–230. ACM, 2004.

[229] Chunyi Peng, Guobin Shen, Yongguang Zhang, Yanlin Li, and Kun Tan. Beepbeep: a high accuracy acoustic ranging system using cots mobile devices. In *Proceedings of the 5th international conference on Embedded networked sensor systems*, pages 1–14. ACM, 2007.

[230] Alex Pentland. Looking at people: Sensing for ubiquitous and wearable computing. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(1):107–119, 2000.

[231] Alex Pentland. Socially aware, computation and communication. *Computer*, 38(3): 33–40, 2005. URL `http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1413116`.

[232] Alex Pentland. Social Signal Processing. *Signal Processing Magazine, IEEE*, 24(4):108–111, 2007. URL `http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4286569`.

[233] Alex Pentland. *Honest signals: how they shape our world*. MIT Press, Cambridge, Mass, 2008. ISBN 0262162563.

[234] Christoph Peppmeier. Speaker diarization for social situation detection. Master's thesis, Technische Universität München; Supervisor: Groh, G., 2011.

[235] Martin Peterson. *An introduction to decision theory*. Cambridge University Press, 2009.

[236] Marcus Pfister and Raul Rojas. Speeding-up backpropagation-a comparison of orthogonal techniques. In *Neural Networks, 1993. IJCNN'93-Nagoya. Proceedings of 1993 International Joint Conference on*, volume 1, pages 517–523, 1993. URL `http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=713967`.

[237] Theophilos Phokas, Hariton Efstathiades, George Pallis, and Marios D. Dikaiakos. Feel the world: A mobile framework for participatory sensing. In *Mobile Web Information Systems*, pages 143–156. Springer, 2013. URL `http://link.springer.com/chapter/10.1007/978-3-642-40276-0_12`.

[238] Fabio Pianesi, Nadia Mana, Alessandro Cappelletti, Bruno Lepri, and Massimo Zancanaro. Multimodal recognition of personality traits in social interactions. In *Proceedings of the 10th international conference on Multimodal interfaces*, pages 53–60. ACM, 2008.

[239] Gerald Pirkl and Paul Lukowicz. Robust, low cost indoor positioning using magnetic resonant coupling. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pages 431–440. ACM, 2012. URL `http://dl.acm.org/citation.cfm?id=2370281`.

[240] Gerald Pirkl and Paul Lukowicz. Resonant magnetic coupling indoor localization system. pages 59–62. ACM Press, 2013. ISBN 9781450322157. doi: 10.1145/2494091.2494108. URL `http://dl.acm.org/citation.cfm?doid=2494091.2494108`.

[241] Lasitha Piyathilaka and Sarath Kodagoda. Gaussian mixture based HMM for human daily activity recognition using 3d skeleton features. In *Industrial Electronics and*

Applications (ICIEA), 2013 8th IEEE Conference on, pages 567–572, 2013. URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6566433.

[242] Ronald L. Plackett. Some theorems in least squares. *Biometrika*, 37(1-2):149–157, 1950. URL http://biomet.oxfordjournals.org/content/37/1-2/149.short.

[243] Isabella Poggi and Francesca D'Errico. Cognitive modelling of human social signals. In *Proceedings of the 2nd international workshop on Social signal processing*, pages 21–26. ACM, 2010. URL http://dl.acm.org/citation.cfm?id=1878124.

[244] Isabella Poggi and Francesca D'Errico. Social Signals: A Psychological Perspective. In *Computer Analysis of Human Behavior*, pages 185–225. Springer London, London, 2011. ISBN 978-0-85729-993-2, 978-0-85729-994-9. URL http://www.springerlink.com/index/10.1007/978-0-85729-994-9_8.

[245] Gerard Pons-Moll, Andreas Baak, Thomas Helten, M. Muller, H.-P. Seidel, and Bodo Rosenhahn. Multisensor-Fusion for 3d Full-Body Human Motion Capture. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 663–670. IEEE, 2010. URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5540153.

[246] Gerard Pons-Moll, Laura Leal-Taixé, Juergen Gall, and Bodo Rosenhahn. Data-driven manifolds for outdoor motion capture. In *Outdoor and Large-Scale Real-World Scene Analysis*, pages 305–328. Springer, 2012. URL http://link.springer.com/chapter/10.1007/978-3-642-34091-8_14.

[247] Gerard Pons-Moll, Jonathan Taylor, Jamie Shotton, Aaron Hertzmann, and Andrew Fitzgibbon. Metric Regression Forests for Human Pose Estimation. 2013. URL http://www.3dtv-con2009.org/papers/data/990/MetricRegressionForests.pdf.

[248] Ronald Poppe. A survey on vision-based human action recognition. *Image and Vision Computing*, 28(6):976–990, June 2010. ISSN 02628856. doi: 10.1016/j.imavis.2009.11.014. URL http://linkinghub.elsevier.com/retrieve/pii/S0262885609002704.

[249] Edy Portmann. The Social Semantic Web. In *The FORA Framework*, pages 13–36. Springer, 2013.

[250] Richard H Price and Dennis L Bouffard. Behavioral appropriateness and situational constraint as dimensions of social behavior. *Journal of Personality and Social Psychology*, 30(4):579, 1974.

[251] Eric A Prigge and Jonathan P How. Signal architecture for a distributed magnetic local positioning system. *Sensors Journal, IEEE*, 4(6):864–873, 2004.

[252] Bodhi Priyantha, Dimitrios Lymberopoulos, and Jie Liu. Littlerock: Enabling energy-efficient continuous sensing on mobile phones. *Pervasive Computing, IEEE*, 10(2):12–15, 2011.

[253] Hairong Qi, S. Sitharama Iyengar, and Krishnendu Chakrabarty. Distributed sensor networks—a review of recent research. *Journal of the Franklin Institute*, 338 (6):655–668, 2001. URL `http://www.sciencedirect.com/science/article/pii/S0016003201000266`.

[254] J Ross Quinlan. *C4.5: Programs for machine learning*. Elsevier, 2014.

[255] Lawrence Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[256] Lawrence R Rabiner and Biing-Hwang Juang. *Fundamentals of speech recognition*, volume 14. PTR Prentice Hall Englewood Cliffs, 1993.

[257] Lawrence R Rabiner, Aaron E Rosenberg, and Stephen E Levinson. Considerations in dynamic time warping algorithms for discrete word recognition. *The Journal of the Acoustical Society of America*, 63(S1):S79–S79, 1978.

[258] Lee Rainie. Cell phone ownership hits 91% of adults, June 2013. URL `http://www.pewresearch.org/fact-tank/2013/06/06/cell-phone-ownership-hits-91-of-adults/`.

[259] Mukund Raj, Sarah H. Creem-Regehr, Kristina M. Rand, Jeanine K. Stefanucci, and William B. Thompson. Kinect Based 3d Object Manipulation on a Desktop Display. In *Proceedings of the ACM Symposium on Applied Perception*, SAP '12, pages 99–102, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1431-2. doi: 10.1145/2338676.2338697. URL `http://doi.acm.org.eaccess.ub.tum.de/10.1145/2338676.2338697`.

[260] William M Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850, 1971.

[261] Cliff Randell, Chris Djiallis, and Henk Muller. Personal position measurement using dead reckoning. In *2012 16th International Symposium on Wearable Computers*, pages 166–166. IEEE Computer Society, 2003. URL `http://www.computer.org/csdl/proceedings/iswc/2003/2034/00/20340166.pdf`.

[262] Jonas Reimers. Detection and modeling of social situations via distances and shoulder-angles, April 2010.

[263] Harry T Reis, John Nezlek, and Ladd Wheeler. Physical attractiveness in social interaction. *Journal of Personality and Social Psychology*, 38(4):604, 1980.

[264] Harry T Reis, Ladd Wheeler, Nancy Spiegel, Michael H Kernis, John Nezlek, and Michael Perri. Physical attractiveness in social interaction: II. Why does appearance affect social experience? *Journal of Personality and Social Psychology*, 43(5):979, 1982.

[265] Martin S Remland, Tricia S Jones, and Heidi Brinkman. Proxemic and haptic behavior in three European countries. *Journal of Nonverbal behavior*, 15(4):215–232, 1991.

[266] Martin S Remland, Tricia S Jones, and Heidi Brinkman. Interpersonal distance, body orientation, and touch: Effects of culture, gender, and age. *The Journal of social psychology*, 135(3):281–297, 1995.

[267] Douglas A. Reynolds and P. Torres-Carrasquillo. The MIT Lincoln Laboratory RT-04f diarization systems: Applications to broadcast audio and telephone conversations. Technical report, DTIC Document, 2004. URL `http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA511688`.

[268] Fazlollah M Reza. *An introduction to information theory*. Courier Dover Publications, 1994.

[269] Virginia P Richmond, James C McCroskey, and Mark Hickson. *Nonverbal behavior in interpersonal relations*. Prentice Hall Englewood Cliffs, NJ, 1991.

[270] Cecilia L Ridgeway and Lynn Smith-Lovin. The gender system and interaction. *Annual review of sociology*, 25(1):191–216, 1999.

[271] Giacomo Rizzolatti and Laila Craighero. The mirror-neuron system. *Annual Review of Neuroscience*, 27(1):169–192, July 2004. ISSN 0147-006X, 1545-4126. doi: 10.1146/annurev.neuro.27.070203.144230. URL `http://www.annualreviews.org/doi/abs/10.1146/annurev.neuro.27.070203.144230`.

[272] Daniel Roetenberg. *Inertial and magnetic sensing of human motion*. PhD thesis, s.n.], S.l., 2006.

[273] Raul Rojas. The Kalman Filter, 2003. URL `http://robocup.mi.fu-berlin.de/buch/kalman.pdf`.

[274] Jeffrey S Rosenschein. *Rules of encounter: designing conventions for automated negotiation among computers*. MIT press, 1994.

[275] Anandarup Roy, Swapan K. Parui, Debyani Nandi, and Utpal Roy. Color image segmentation using a semi-wrapped gaussian mixture model. In *Pattern Recognition and Machine Intelligence*, pages 148–153. Springer, 2011. URL `http://link.springer.com/chapter/10.1007/978-3-642-21786-9_26`.

[276] A.M. Sabatini. Quaternion-Based Extended Kalman Filter for Determining Orientation by Inertial and Magnetic Sensing. *IEEE Transactions on Biomedical Engineering*, 53(7):1346–1356, July 2006. ISSN 0018-9294. doi: 10.1109/TBME.2006.875664. URL `http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1643403`.

[277] Debashis Saha and Amitava Mukherjee. Pervasive computing: a paradigm for the 21st century. *Computer*, 36(3):25–31, 2003. URL `http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1185214`.

[278] Albert Ali Salah, Maja Pantic, and Alessandro Vinciarelli. Recent developments in social signal processing. In *Systems, Man, and Cybernetics (SMC), 2011 IEEE International Conference on*, pages 380–385. IEEE, 2011. URL `http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6083695`.

[279] Stan Salvador and Philip Chan. FastDTW: Toward accurate dynamic time warping in linear time and space. *Intelligent Data Analysis*, 11(5):561–580, 2007.

[280] T. Saponas, Jonathan Lester, Jon Froehlich, James Fogarty, and James Landay. iLearn on the iPhone: Real-Time Human Activity Classification on Commodity Mobile Phones. *University of Washington CSE Tech Report UW-CSE-08-04-02*, 2008, 2008. URL `http://research.microsoft.com/en-us/um/people/ssaponas/publications/UW-CSE-08-04-02.pdf`.

[281] Mahadev Satyanarayanan. Pervasive computing: Vision and challenges. *Personal Communications, IEEE*, 8(4):10–17, 2001. URL `http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=943998`.

[282] Albert E Scheflen. Communication and regulation in psychotherapy. *Psychiatry: Journal for the Study of Interpersonal Processes*, 1963.

[283] Albert E Scheflen. *Communicational structure: Analysis of a psychotherapy transaction.* Indiana U. Press, 1973.

[284] Albrecht Schmidt and Kristof Van Laerhoven. How to build smart appliances? *Personal Communications, IEEE*, 8(4):66–71, 2001. URL `http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=944006`.

[285] Albrecht Schmidt, KofiAsante Aidoo, Antti Takaluoma, Urpo Tuomela, Kristof Laerhoven, and Walter Velde. Advanced Interaction in Context. In Hans-W. Gellersen, editor, *Handheld and Ubiquitous Computing*, volume 1707 of *Lecture Notes in Computer Science*, pages 89–101. Springer Berlin Heidelberg, 1999. ISBN 978-3-540-66550-2. URL `http://dx.doi.org/10.1007/3-540-48157-5_10`.

[286] Tobias Schröder, Janine Netzel, Carsten C Schermuly, and Wolfgang Scholl. Culture-constrained affective consistency of interpersonal behavior: A test of affect control theory with nonverbal expressions. *Social Psychology*, 44(1):47, 2013.

[287] B.W. Schuller. *Intelligent Audio Analysis*. Signals and Communication Technology. Springer, 2013. ISBN 9783642368066. URL `http://books.google.de/books?id=kEm9BAAAQBAJ`.

[288] Rick Schwartz. Early Access to Snapdragon 805 – Develop for Tomorrow's Devices Today, April 2014. URL `https://developer.qualcomm.com/blog/early-access-snapdragon-805-develop-tomorrow-s-devices-today`.

[289] Loren Schwarz. *Machine Learning for Human Motion Analysis and Gesture Recognition.* PhD thesis, Technische Universitaet Muenchen, 2012.

[290] Kari Sentz and Scott Ferson. *Combination of evidence in Dempster-Shafer theory*, volume 4015. Citeseer, 2002.

[291] Francesco Setti, Oswald Lanz, Roberta Ferrario, Vittorio Murino, and Marco Cristani. Multi-Scale F-Formation Discovery for Group Detection. *ICIP*, 2013. URL `http://www.ieeeicip.org/Proc/pdfs/0003547.pdf`.

[292] Francesco Setti, Chris Russell, Chiara Bassetti, and Marco Cristani. F-formation Detection: Individuating Free-standing Conversational Groups in Images. *arXiv preprint arXiv:1409.2702*, 2014.

[293] Lawrence J Severy, Donelson R Forsyth, and Peggy Jo Wagner. A multimethod assessment of personal space development in female and male, black and white children. *Journal of Nonverbal Behavior*, 4(2):68–86, 1979.

[294] Glenn Shafer. *A mathematical theory of evidence.* Princeton University Press, 1976. ISBN 978-0691100425.

[295] Yuanchun Shi. Context awareness, the spirit of pervasive computing. In *Pervasive Computing and Applications, 2006 1st International Symposium on*, pages 6–6. IEEE, 2006.

[296] Yue Shi, Yuanchun Shi, and Jie Liu. A rotation based method for detecting on-body positions of mobile devices. In *Proceedings of the 13th international conference on Ubiquitous computing*, pages 559–560, 2011. URL `http://dl.acm.org/citation.cfm?id=2030212`.

[297] Ken Shoemake. Animating rotation with quaternion curves. *ACM SIGGRAPH computer graphics*, 19(3):245–254, 1985. URL `http://dl.acm.org/citation.cfm?id=325242`.

[298] Jamie Shotton, Toby Sharp, Alex Kipman, Andrew Fitzgibbon, Mark Finocchio, Andrew Blake, Mat Cook, and Richard Moore. Real-time Human Pose Recognition in Parts from Single Depth Images. *Commun. ACM*, 56(1):116–124, January 2013. ISSN 0001-0782. doi: 10.1145/2398356.2398381. URL `http://doi.acm.org.eaccess.ub.tum.de/10.1145/2398356.2398381`.

[299] S Shozo. *Comfortable distance between people: Personal Space.* Japan Broadcast Publishing Co., Ltd, 1990.

[300] Malcolm D Shuster. Deterministic three-axis attitude determination. *Journal of Astronautical Sciences*, 52(3):405–419, 2004.

[301] Robert Shuter. A field study of nonverbal communication in Germany, Italy, and the United States. *Communications Monographs*, 44(4):298–305, 1977.

[302] Matthew A. Siegler, Uday Jain, Bhiksha Raj, and Richard M. Stern. Automatic segmentation, classification and clustering of broadcast news audio. In *Proc. DARPA Broadcast News Workshop*, page 11, 1997. URL `http://www.cs.cmu.edu/afs/cs/user/msiegler/www/publish/DARPA/darpa97-H4seg.pdf`.

[303] John R. Smith, Milind Naphade, and Apostol Natsev. Multimedia semantic indexing using model vectors. In *Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 International Conference on*, volume 2, pages II–445. IEEE, 2003. URL `http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1221649`.

[304] Steven W. Smith. *Digital Signal Processing: A Practical Guide for Engineers and Scientists*. Newnes, 2003. URL `http://books.google.com/books?hl=en&lr=&id=PCrcintuzAgC&oi=fnd&pg=PP1&dq=%22available.+Computers+were+expensive+during+this+era,+and+DSP%22+%22who+thought+they+could+make+money+in+the+rapidly+expanding+field%22+%22sensory+analysis%22+%22of+these+diverse+applications+are+shown%22+&ots=TsLYggXJfM&sig=JKLhN-XxYh_XG89BNPEqZWUxhl0`.

[305] Robert Sommer. Leadership and Group Geometry. *Sociometry*, (24):99–100, 1961.

[306] Robert Sommer. Further studies of small group ecology. *Sociometry*, 1965.

[307] Robert Sommer. *Studies in personal space*. Bobbs-Merrill, 1967.

[308] Robert Sommer. *Personal Space: The Behavioral Basis of Design*. Prentice Hall, 1969. ISBN 978-0136575771.

[309] Robert Sommer. Personal space in a digital age. *Handbook of environmental psychology*, pages 647–660, 2002.

[310] C. Song, Z. Qu, N. Blumm, and A.-L. Barabasi. Limits of Predictability in Human Mobility. *Science*, 327(5968):1018–1021, February 2010. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1177170. URL `http://www.sciencemag.org/cgi/doi/10.1126/science.1177170`.

[311] Bonnie G. Steele, Basia Belza, Kevin Cain, Catherine Warms, Jeff Coppersmith, and J. Howard. Bodies in motion: monitoring daily activity and exercise with motion sensors in people with chronic pulmonary disease. *Journal of rehabilitation research and development*, 40(5; SUPP/2):45–58, 2003. URL `http://www.rehab.research.va.gov/jour/03/40/5sup2/pdf/steele.pdf`.

[312] Ulrich Steinhoff and Bernt Schiele. Dead reckoning from the pocket-an experimental study. In *Pervasive Computing and Communications (PerCom), 2010 IEEE International Conference on*, pages 162–170. IEEE, 2010. URL `http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5466978`.

[313] William Storms, Jeremiah Shockley, and John Raquet. Magnetic field navigation in an indoor environment. In *Ubiquitous Positioning Indoor Navigation and Location Based Service (UPINLBS), 2010*, pages 1–10. IEEE, 2010. URL `http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5653681`.

[314] John Stowers, Michael Hayes, and Andrew Bainbridge-Smith. Altitude control of a quadrotor helicopter using depth map from Microsoft Kinect sensor. In *Mechatronics (ICM), 2011 IEEE International Conference on*, pages 358–362. IEEE, 2011.

[315] Eric Sundstrom and Irwin Altman. Interpersonal relationships and personal space: Research review and theoretical model. *Human Ecology*, 4(1):47–67, 1976.

[316] Nan M Sussman and Howard M Rosenfeld. Influence of culture, language, and sex on conversational distance. *Journal of Personality and Social Psychology*, 42(1):66, 1982.

[317] S. P. Tarzia, P. A. Dinda, R. P. Dick, and G. Memik. Indoor localization without infrastructure using the acoustic background spectrum. In *Proceedings of the 9th international conference on Mobile systems, applications, and services*, pages 155–168, 2011. URL `http://dl.acm.org/citation.cfm?id=2000011`.

[318] Antonio Terracciano, Robert R McCrae, Larry J Brant, and Paul T Costa Jr. Hierarchical linear modeling analyses of the NEO-PI-R scales in the Baltimore Longitudinal Study of Aging. *Psychology and aging*, 20(3):493, 2005.

[319] Paul Thagard. *Coherence in thought and action.* MIT Press, 2002.

[320] Henri Theil. On the estimation of relationships involving qualitative variables. *American Journal of Sociology*, pages 103–154, 1970.

[321] Joy A Thomas and Joy A Thomas. *Elements of information theory.* Wiley New York, 2006.

[322] Marc Thylmann. Presseinformation zur BITKOM/Aris Umfrageforschung 2013, March 2013. URL `http://www.bitkom.org/de/markt_statistik/64046_77178.aspx`.

[323] David H. Titterton, John Weston, et al. *Strapdown Inertial Navigation Technology.* Institution of Engineering & Technology, 2004.

[324] S.E. Tranter and D.A. Reynolds. An overview of automatic speaker diarization systems. *IEEE Transactions on Audio, Speech and Language Processing*, 14(5):1557–1565, September 2006. ISSN 1558-7916. doi: 10.1109/TASL.2006.878256. URL `http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1677976`.

[325] Harry C Triandis. The self and social behavior in differing cultural contexts. *Psychological review*, 96(3):506, 1989.

[326] Alain Tritschler and Ramesh Gopinath. Improved speaker segmentation and segments clustering using the Bayesian information criterion. In *Proc. Eurospeech*, volume 2, pages 679–682, 1999. URL `http://www.research.ibm.com/people/r/rameshg/tritschler-eurospeech99.pdf`.

[327] P. Turaga, R. Chellappa, V.S. Subrahmanian, and O. Udrea. Machine Recognition of Human Activities: A Survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(11):1473–1488, November 2008. ISSN 1051-8215, 1558-2205. doi: 10.1109/TCSVT.2008.2005594. URL `http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4633644`.

[328] Matthew Turk and Alex Pentland. Eigenfaces for recognition. *Journal of cognitive neuroscience*, 3(1):71–86, 1991.

[329] Matthew A Turk and Alex P Pentland. Face recognition using eigenfaces. In *Computer Vision and Pattern Recognition, 1991. Proceedings CVPR'91., IEEE Computer Society Conference on*, pages 586–591. IEEE, 1991.

[330] David Uzzell and Nathalie Horne. The influence of biological sex, sexuality and gender role on interpersonal distance. *British Journal of Social Psychology*, 45(3): 579–597, 2006.

[331] Alireza Vahdatpour, Navid Amini, and Majid Sarrafzadeh. On-body device localization for health and medical monitoring applications. In *Pervasive Computing and Communications (PerCom), 2011 IEEE International Conference on*, pages 37–44. IEEE, 2011. URL `http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5767593`.

[332] Russell Veitch, Andrew Getsinger, and Daniel Arkkelin. A note on the reliability and validity of the Comfortable Interpersonal Distance Scale. *The Journal of Psychology*, 94(2):163–165, 1976.

[333] A. Vinciarelli, M. Pantic, and H. Bourlard. Social signal processing: Survey of an emerging domain. *Image and Vision Computing*, 27(12):1743–1759, 2009. URL `http://www.sciencedirect.com/science/article/pii/S0262885608002485`.

[334] Alessandro Vinciarelli, Maja Pantic, Hervé Bourlard, and Alex Pentland. Social signals, their function, and automatic analysis: a survey. In *Proceedings of the 10th international conference on Multimodal interfaces*, pages 61–68, 2008. URL `http://dl.acm.org/citation.cfm?id=1452405`.

[335] Alessandro Vinciarelli, Roderick Murray-Smith, and Hervé Bourlard. Mobile social signal processing: vision and research issues. In *Proceedings of the 12th international*

conference on Human computer interaction with mobile devices and services, pages 513–516, 2010. URL `http://dl.acm.org/citation.cfm?id=1851731`.

[336] Alessandro Vinciarelli, M. Pantic, D. Heylen, C. Pelachaud, I. Poggi, F. D'Errico, and M. Schroeder. Bridging the Gap between Social Animal and Unsocial Machine: A Survey of Social Signal Processing. *IEEE Transactions on Affective Computing*, 3(1):69–87, 2012. ISSN 1949-3045. doi: 10.1109/T-AFFC.2011.27. URL `http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5989788`.

[337] Paul Viola and Michael Jones. Robust real-time object detection. *International Journal of Computer Vision*, 4:34–47, 2001.

[338] Silke Wagner and Dorothea Wagner. *Comparing clusterings: an overview.* Universität Karlsruhe, Fakultät für Informatik Karlsruhe, 2007.

[339] O. Michael Watson. On Proxemic Research. *Current Anthropology*, 10(2):3, 1969.

[340] O. Michael Watson and Theodore D. Graves. Quantitative Research in Proxemic Behavior. *American Anthropologist*, 68(4):971–985, 1966.

[341] Duncan J Watts and Steven H Strogatz. Collective dynamics of "small-world" networks. *nature*, 393(6684):440–442, 1998.

[342] Greg Welch and Gary Bishop. *An introduction to the Kalman filter.* 1995.

[343] Gregory J Welk, Jerome A Differding, Raymond W Thompson, Steven N Blair, Jim Dziura, and Peter Hart. The utility of the Digi-walker step counter to assess daily physical activity patterns. *Medicine and science in sports and exercise*, 32(9 Suppl): S481–8, 2000.

[344] Jan Wendel. *Integrierte Navigationssysteme: Sensordatenfusion, GPS und Inertiale Navigation.* Oldenbourg Verlag, 2007.

[345] Chung-Yi Weng, Wei-Ta Chu, and Ja-Ling Wu. RoleNet: Movie Analysis from the Perspective of Social Networks. *IEEE Transactions on Multimedia*, 11(2):256–271, February 2009. ISSN 1520-9210. doi: 10.1109/TMM.2008.2009684. URL `http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4757440`.

[346] Candace West and Don H Zimmerman. Doing gender. *Gender & society*, 1(2): 125–151, 1987.

[347] Widyawan, Gerald Pirkl, Daniele Munaretto, Carl Fischer, Chunlei An, Paul Lukowicz, Martin Klepal, Andreas Timm-Giel, Joerg Widmer, Dirk Pesch, and Hans Gellersen. Virtual lifeline: Multimodal sensor data fusion for robust navigation in unknown environments. *Pervasive and Mobile Computing*, 8(3):388–401, June 2012. ISSN 15741192. doi: 10.1016/j.pmcj.2011.04.005. URL `http://linkinghub.elsevier.com/retrieve/pii/S1574119211000411`.

[348] Ian H. Witten and Eibe Frank. *Data Mining: Practical Machine Learning Tools and Techniques.* Morgan Kaufmann, 2005.

[349] Oliver J. Woodman. An introduction to inertial navigation. *University of Cambridge, Computer Laboratory, Tech. Rep. UCAMCL-TR-696*, 14:15, 2007. URL `http://narwhalroboticsquad.googlecode.com/svn/trunk/Quad/Kalman%20zips/10.1.1.63.7402.pdf`.

[350] Wendong Xiao, Wei Ni, and Yue Khing Toh. Integrated Wi-Fi fingerprinting and inertial sensing for indoor positioning. In *Indoor Positioning and Indoor Navigation (IPIN), 2011 International Conference on*, pages 1–6. IEEE, 2011. URL `http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6071921`.

[351] Lei Xu and Michael I Jordan. On convergence properties of the EM algorithm for Gaussian mixtures. *Neural computation*, 8(1):129–151, 1996.

[352] Ronald R Yager. On the Dempster-Shafer framework and new combination rules. *Information sciences*, 41(2):93–137, 1987.

[353] Ching-man Au Yeung, Ilaria Liccardi, Kanghao Lu, Oshani Seneviratne, and Tim Berners-Lee. Decentralization: The future of online social networking. In *W3C Workshop on the Future of Social Networking Position Papers*, volume 2, pages 2–7, 2009.

[354] Lotfi A. Zadeh. A simple view of the Dempster-Shafer theory of evidence and its implication for the rule of combination. *AI magazine*, 7(2):85, 1986. URL `http://www.aaai.org/ojs/index.php/aimagazine/article/viewArticle/542`.

[355] Beibei Zhan, Dorothy N. Monekosso, Paolo Remagnino, Sergio A. Velastin, and Li-Qun Xu. Crowd analysis: a survey. *Machine Vision and Applications*, 19(5-6):345–357, October 2008. ISSN 0932-8092, 1432-1769. doi: 10.1007/s00138-008-0132-4. URL `http://link.springer.com/10.1007/s00138-008-0132-4`.

[356] Li Zhang, Brian Curless, and Steven M. Seitz. Rapid shape acquisition using color structured light and multi-pass dynamic programming. In *3D Data Processing Visualization and Transmission, 2002. Proceedings. First International Symposium on*, pages 24–36. IEEE, 2002. URL `http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1024035`.

[357] Zhengyou Zhang. Microsoft Kinect Sensor and Its Effect. *MultiMedia, IEEE*, 19(2):4–10, 2012. URL `http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6190806`.

[358] Bowen Zhou and John Hansen. Unsupervised Audio Stream Segmentation And Clustering Via The Bayesian Information Criterion. In *in Proc. ISCLP 2000*, pages 714–717, 2000.

[359] Udo Zölzer, Xavier Amatriain, and John Wiley. *DAFX: Digital Audio Effects*, volume 1. Wiley Online Library, 2nd edition, 2011. ISBN 978-0-470-66599-2. URL `http://onlinelibrary.wiley.com/doi/10.1002/9781119991298.fmatter/summary`.