PROMETHEUS DATABASE: A MULTIMODAL CORPUS FOR RESEARCH ON MODELING AND INTERPRETING HUMAN BEHAVIOR

Stavros Ntalampiras[1], Dejan Arsić[2], Andre Störmer[2], Todor Ganchev[1], Ilyas Potamitis[3] and Nikos Fakotakis[1]

[1]Department of Electrical and Computer Engineering, University of Patras,
[2]Institute for Human Machine Communication, Technische Universität München,
[3]Technological Educational Institute of Crete

ABSTRACT

The present paper describes the construction of a multimodal database, referred to as the PROMETHEUS database, which contains recordings from heterogeneous sensors. The main purpose of this database is the development of a framework for monitoring and interpretation of human behavior in unrestricted environments of both indoor and outdoor type. It contains single-person and multi-person scenarios, but also covers scenarios with interactions between groups of people. It is devoted to detection of typical and atypical events, while care has been to taken for the recordings to be as close to real-world conditions as possible. The uniqueness of the PROMETHEUS database comes not only from the unique sensor sets but is due primarily to its generic design, which allows for embracing a wide range of real-world applications (including smart-home and human-robot interaction interfaces, indoors/outdoors public areas surveillance etc).

Index Terms— Multimodal database, heterogeneous sensors, signal-based surveillance, civil safety.

1. INTRODUCTION

Prediction and interpretation of human behavior using probabilistic structures and heterogeneous sensors (PROMETHEUS) is a project funded under the umbrella of EC-FP7 and proposes to research probabilistic inference algorithms within the paradigm of recursive Bayesian estimation to the problem of online tracking of multiple people in a scene, and the identification of the interaction amongst them and with the environment. The core research components of the proposal are in the representation of uncertainties arising from multiple modalities including laser, visual, acoustical and infrared, the fusion of information gathered from such diverse range of sensors into a coherent mechanism that makes predictions about interactions and the coupling between modeling high level behavior of people in a scene with signal processing issues of sensor fusion and tracking. Bayesian inference will provide the core architectural framework to carry out the above research via a rigorous mathematical framework. Sequential estimation algorithms around the notion of particle filtering and approximations that are required to handle estimation variance issues will be explored. The overall task is based on a hierarchy which at low levels consists of processing of robust detection of humans, and the expansion of this into abstractions as 2D shapes and 3D wholes of articulated limbs. The subsequent levels are responsible for the localization and tracking of humans on an extended time basis. Higher level processing includes making predictions about intended actions of humans in the scene via a combination of data-driven (i.e., classifications learnt from an annotated database), continuous time (vector autoregressive processes) and utility-driven models.

Specifically, the PROMETHEUS project will:

- deploy a dispersed network of perceptual modalities that will allow an integration of sensing views and efficient analysis of human-motion analysis, including situations with several people in the scene
- set-up a synergetic network of heterogeneous modalities that provide complementary perceptual information
- implement a probabilistic framework for the fusion of heterogeneous information to allow the detection, localization and tracking of humans
- train advanced probabilistic networks to recognize human behavioral and interaction patterns based on sensor-logs of the heterogeneous sensors
- provide short-time prediction of human goals based on the trained models of behavior

The present paper analyzes in detail the construction of the multisensor database that was implemented during the project and enables the training and evaluation of the probabilistic structures that identify, track and recognize human actions.

978-1-4244-3298-1/09/$25.00 ©2009 IEEE

2. APPLICATION SCENARIOS

The indoor and outdoor scenarios outlined in the following subsections serve to illustrate the wide range of applications that the technology developed within the PROMETHEUS project enables. The indoors scenarios include smart-home environment and public area, while the outdoor scenarios incorporate surveillance applications at airports, ATMs, and left luggage scenarios. Likewise these scenarios can be transferred to other security relevant settings, as these included general suspicious behavior patterns.

2.1. Indoors

Human-friendly multimodal interaction interface for the needs of smart-home environments is among the most interesting applications that arise from the PROMETHEUS technology. In brief, smart-home applications aim at improved homecare for elderly people living on their own, improved healthcare for chronic patients, improved quality of live of people with motor disabilities, etc. However, here we do not aim at illustrating how the needs of the aforementioned user groups are satisfied, but instead at illustrating the underlying functionality that is common in the different setups, and that the PROMETHEUS database implements. Specifically, the smart-home environment scenarios covered in the database contain five single-person and fourteen multiple-person action scripts. These basic scenarios implement: (i) typical examples of human-friendly interaction with a virtual home assistant, which controls the smart-home appliances, and (ii) sequences of actions in support of research on human behavior monitoring and interpretation. Three of the single-person and eleven of the multiple-person scenarios involve situations, where atypical events or actions, such as: person falling or lying on the floor, fire alarm followed by panic, dropping/breaking objects, etc, occurred. Each of the nineteen scenarios was recorded between three and five times, which differed by the actors involved in the role-playing and by the actual interpretation of the written scripts.

In this case the actors were all professionals with relatively high experience. The family visit case was the one where five actors executed the respective scenario. That was a typical situation where the system should recognize that nothing abnormal is going on but when one of them fainted and/or another atypical event happened it should detect the abnormality and especially what kind of danger the people were in. A decision on what are the appropriate actions that should be taken by the system comprises the higher level of scene analysis

The second indoors setting of interest aimed at capturing data which are representative of the manner in which people move and behave in public areas. In this scenario overlapping and non-overlapping views showing narrow aisles were recorded. People were moving in a narrow environment, leaving and entering through different field of views. Occlusion was typical if more than one person was moving due to space limitations. For illustration in Fig. 1 we show two different narrow aisles, one entrance area with a larger door, and one open area.

This scenario has been chosen in order to aid the development of re-identification algorithms that enable to re-identify tracks of persons or objects that enter the scene and have previously been seen. This can be either in the same view or a re-identification of tracks between different views. For this purpose, two scenes have been captured, where the actors enter the scene, stand still on a predefined place and walk along a predefined route while the actions were performed in different places. After that, a different and more challenging scenario has been recorded, which was unconstrained and included one or more than one actors moving at the same time. It should be mentioned that people often carried bags and/or interacted with other people, including speech, gestures, fighting, tumbling and theft of baggage items. The re-identification task is of major importance besides the behavioral analysis. Biometric patterns, such as faces, cannot be used, as these are not visible all the time due to inconvenient fields of view or severe occlusion. These problems are usually encountered under real world conditions and therefore other characteristic patterns have to be identified.

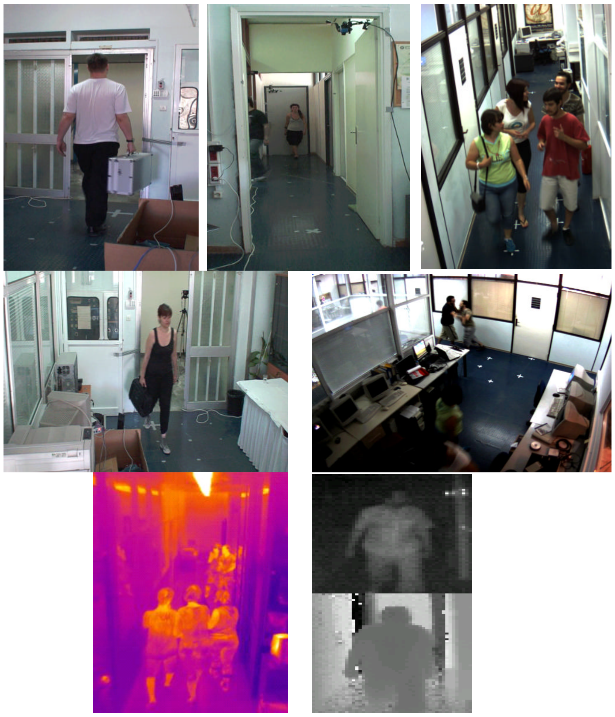

Fig. 1. Captured images of different narrow aisles, one entrance area with a larger door and one open area. Sensors used include video (top 5 images), thermal infrared (lower left) and a PMD device (lower right).

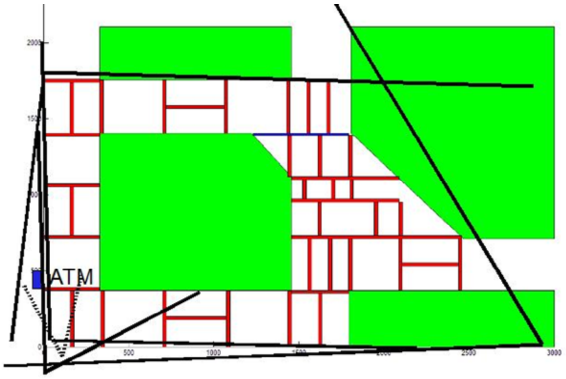

Fig. 2. Position of video sensors and their field of view in the outdoors scenarios (solid lines). The dotted line indicates the position of the PMD camera during the recording of the ATM scenario.

2.1. Outdoors

European authorities have shown interest in the automated detection of security related events in the recent years. These are widely independent of environmental influences and therefore the scenarios could also be recorded in an outdoor site. This includes the additional challenge of difficult and changing lighting situations, moving background, such as trees and bushes, and various floor patterns.

For all scenarios there were no fixed rules for timing and spatial relationship, such as in the PETS2007 challenge [1]. The actors were instructed to perform a defined activity and improvise. This way some variance was included into the dataset, avoiding redundancy and providing a more challenging recognition task, as simple heuristics [2] should not work. The outdoor recordings were conducted at the previously described site and split into two sessions, which implemented the security related scenarios: *ATM* and *airport*. These contained suspicious events at an ATM and an extended set of luggage related events [3]. All visual sensors were setup in overlapping views, with three high resolution cameras mounted on top of the building, and this way providing an overview over the scene. An additional PAL camera, a stereo camera and the PMD sensor were placed as detail sensors near the simulated ATM to obtain a detailed view of the persons operating and standing near the ATM machine. A microphone array has been placed behind the ATM to pick up acoustic sound events (vocalic and/or non-vocalic ones). The sensor arrangement is shown in Fig. 2. Note the large area that is covered by the overview sensors, in order to provide a large recording site with various obstacles and view angles. Note that the bushes located in the scenario are not simply obstacles but are really challenging objects to handle, as these were constantly moving in the wind and created very dark shadows were pedestrians could not be seen by the camera.

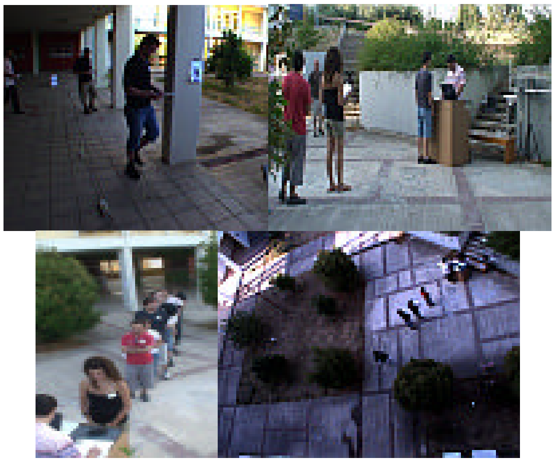

Fig. 3. Descriptive images of the airport scenario, with detailed views at the counter and the non-overlapping corridor area where enrollment is performed.

Initially, the actors were told to normally walk along the path and lawn randomly in the beginning and occasionally leave the site. Some of them were carrying luggage with them or just approaching the ATM operating it and get some cash. To complicate the tracking task they were meeting friends and greeting each other, creating a merger in the visual domain. In the second phase ATM theft scenarios were introduced, where one person would loiter in a specific area and observe another one operating the ATM. As soon as the machine releases the money, the thief approaches the ATM, grabs the money and runs away. The robbed person now either follows the thief or stays in the same place, as she was not able to run, which would be the case for elderly people. In both cases abnormal vocalic manifestations (e.g. screams) for help have been captured. While the ATM scenario has been recorded other actors were conducting luggage related scenarios. These included leaving luggage behind at a random location, swapping luggage with another person and a person stealing another person's suitcase. All these scenarios are supposed to be security relevant characteristic of typical and atypical situations and can be simulated easily.

The airport scenario addresses several problems that are concerning the authorities at the moment. One of the major issues is tracking passengers even through non-overlapping camera views or re-recognition after leaving sensitive areas, such as lavatories, where cameras are not installed. Therefore the recording site has been extended to a building bordering the open-air site, where the actors had to pass. This area was used for the initialization of a person's ID, which has to be recognized later on. After leaving the building they had to approach a counter, representing an airport's passport control point, and stand in line if necessary. To create realistic scenarios both individuals and groups were walking along the assigned path, as these are predefined at most international airports. While waiting in line at the counter *interesting* events, related to airport security, were happening: (i) an argument at the counter, (ii) people fighting, while standing in line, (iii) person begging for money, and (iv) group of people gathering around a celebrity. Besides the overview cameras used in the ATM scenario several other cameras have been installed to pick up more details of the activities happening around the counter. Additionally a thermal infrared camera has been recording the scenario from top of the building in order to resolve illumination issues.

3. CONTENTS OF THE DATABASE

The database design was based on the requirements of the application scenarios outlined in Section 2 and on the requirements evolving from the use of probabilistic framework for data processing and fusion. Specifically, the application scenarios defined the choice of test sites, which simulate the target environments, the contents of the action scripts, the number and contents of the task cards, the number of actors, the interactions between the actors, the interactions between actors and objects, etc. On the other hand, the technology requirements set the margins for the number of implementations of each script, the length of the individual recording sessions, the total size of the database, and least but not last the choice of sensor set for each setup.

Eventually we came up with a database design which had a more general nature than the strict requirements of the target applications and covers a wide range of indoors and outdoors scenarios. The completed database comprised of a number of recording sessions, which implement different aspects of the given application scenarios. All recordings belonging to a single session shared a common equipment setup, and represented a number of controlled conditions (for instance, in the smart-home application scenario one session was devoted to single-person actions, another for multiple-person interactions, etc). Each session was comprised of multiple action scenes concatenated in a single sequence, where each action scene was implemented a number of times, with different actors and different objects. The length of various sessions varied between 15 and 60 minutes, and the sensor set involved varied between 10 and 14 heterogeneous sensors.

Six sessions were recorded in order to fulfill the requirements of the airport scenario with average duration 21 minutes. Two sessions of 30 and 60 minutes duration respectively were found sufficient for the ATM - security scenario. Regarding the smart-home scenario three sessions with 22 minutes of average duration were captured.

3.1. Actors and Task Cards

Five professional actors and dozen of supernumerary actors were needed for implementation of the different scenarios:
(i) the *smart-home scenario* was entirely implemented by five professional actors with significant experience.
(ii) *ATM scenario*: Ten actors were needed for the purposes of this scenario. Four of them were professional ones and carried out major part in the performance of the abnormal situations that happened during this scenario. Atypical situations were played many times with great variations between them so as to obtain representative data. Acts like luggage swap, left luggage, theft etc were recorded and executed by more than three actors each.
(iii) *Airport scenario*: Fifteen actors were recruited in this case, five of which were professionals. A great deal of actions needed to be recorded for the needs of this scenario so each action was performed by more than four persons (especially the atypical ones). Actions like person loitering, a bag being left unattended by a person who is leaving, a bag picked up by a person other than the one bringing it in, a celebrity attracting attention and people following, a person threatening others or molesting others etc were caught with great variations. It should be mentioned that the recordings of the airport scenario took place at both indoor and outdoor environments.

Actors' actions were guided by task cards. Particular care was taken for explaining a-priori the background and purpose of filming in terms of what are the general expectations of the project as well as each of the scenes separately. This procedure helped both professional and non-professional actors who took part in the recordings. The task cards were divided into two main categories: one person and multiple persons scenarios. The description of the actions to be performed were written in a clear and complete way while they cover a wide range of activities that represent real-life situations. The actors were instructed to perform normal everyday actions as well as abnormal ones which are indicative of catastrophic events. A specific template was followed which includes the number of the respective action as well as the action itself written as brief as possible with emphasis on being understandable by the side of the person who is going to perform the action. The actions were performed at least one time following the guidelines given on the task cards but also the actors were given the chance to perform the scenario on their own and improvise so as to capture as realistic behaviors as possible.

3.2. Sessions and Environmental Conditions

The data that comprise the PROMETHEUS project database has been captured during two recording campaigns each of which lasted two days and took place in June and July 2008, respectively. The first one was mainly comprised of calibrating and testing the equipment while most of the data were recorded at the second campaign. Initially the positions of all the heterogeneous sensors including video cameras, thermal sensors, and microphone arrays were decided. Cameras were distributed across the recording sites so as to obtain both close and global view of the actors as well as all *interesting* events. Microphone arrays were placed close to where most of the action was to take place while care has been taken not to enter the field of view of any camera. The best sensor configuration in terms of capabilities regarding the area coverage was determined during the initial recording campaign.

Emphasis was placed upon recording the scenarios while having a high degree of variation in terms of the following sources: (i) human appearance: actors changed clothes and accessories, (ii) viewpoint: information is to be fused from heterogeneous modalities and the scenarios included both normal and abnormal situations captured from dissimilar viewpoints, (iii) illumination: shadows and illumination variations should not confuse our system and recording were carried out in the morning, at mid-day, in the afternoon and at night (which advanced the thermal cameras because of the high temperatures during daytime) and (iv) background: the setting was altered on a regular basis using random objects. All the former parameters make automatic tracking and recognition of people, situations and activities very challenging tasks in practical applications. The different influencing variables were included in the database so that they can either be modeled or for creating algorithms which tolerate such distortions.

3.4. Sensors

CCTV systems currently use standard video cameras for data acquisition, as these are widely available and also affordable. In order to get the most out of the video each sensor's properties had to be chosen carefully. For detail views located near the scenery cameras with standard PAL resolution are used, as this detail level is considered sufficient for small distances. For capturing wide spaces the distance between sensors and objects has to be enlarged. A high level of detail is provided the use of high resolution cameras, where 1024x768 pixels are considered as minimum. This can guarantee a minimum size of objects of interest in the visual data. To receive a large field of view optics with short focal length have been chosen, providing a wide angled view on the cost of radial distortion, which can be removed in the subsequent calibration step.

Traditional NIR cameras usually emit invisible infrared light to illuminate the scenery and are hence independent of the lighting situation. This technique is not applicable in every scenario, e.g. warfare, as the emitted light can be traced and affect other sensors or technical equipment.

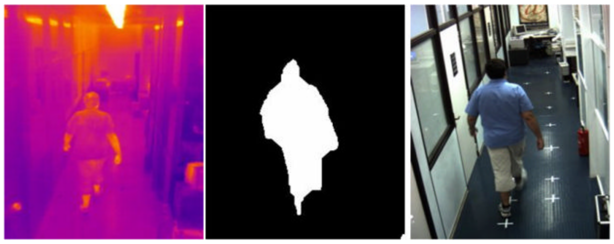

Fig. 4 Exemplary Thermal Image and extracted foreground.

Therefore thermography is frequently applied. Each object in the real world with a temperature above absolute zero emits light in an invisible spectrum, as its wavelength is too large. The higher the object's temperature, the greater the IR radiation emitted. Infrared thermography cameras produce images of invisible infrared or "heat".

A very convenient side effect is the ability to handle thermal infrared data the same way as standard imagery. Thus e.g. foreground segmentation techniques, such as GMM, can be applied without any changes (see Fig. 4). Especially CCD sensors have one drawback in surveillance applications. The distance to camera and other objects cannot be determined exactly, which is rather important for the description of activities and interaction. Thus it has been decided to install 3D sensors in sensitive regions. A popular sensor, unfortunately with limited reach, is the widely used Photonic Mixture Device [4].

The image and depth information acquisition principle of the Photonic PMD is based on the run-time difference of a light impulse directly send to the detector and the reflected light from the surface of objects in the environment. In Fig. 5 the simplified so-called time-of-flight measurement principle for smart pixel is shown. With utmost precise counters, emitters and receivers the distance between the camera pixel and the object can be approximated by $d = t/2*C$, where t represents the measured turnaround time between the start of a light impulse and its return to the receiver. The variable C represents the speed of light. The measurement of the flight-time is carried out using the phase shift of modulated infrared light pulses. By combining several smart pixels in a two dimensional structure an image sensor with fully parallel operating cells arises, allowing the 3D surface reconstruction of the scene. Since this measurement paradigm is directly implemented in the detector's hardware there is no additional computational effort, such as that arising from stereo

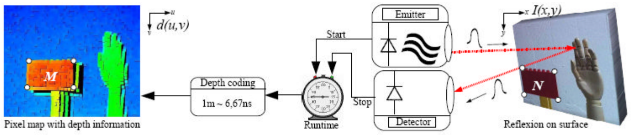

Fig. 5 Time-of-flight measurement principle and calibration body from both sensors' perspectives [6].

cameras. The refresh rate for one measurement loop allows between 5 and 50 frames/second.

To overcome the problem of background illumination, which superposes the running pulse, various further techniques, such as optical filters and active circuits are implemented on the chip. The sensor's usage of the suppression of background illumination makes it even possible to suppress the effects of bright ambient light thus this measurement becomes independent from existing lighting conditions. The emitted infrared light has a wavelength of 870nm. By integrating the received light impulses over a certain interval the PMD camera could further serve as a NIR infrared camera.

However, there are still some drawbacks that the employed measurement technique is suffering from. Range measurement problems occur in conjunction with highly reflective surfaces that are too close to the sensor. By the mirroring effect of the infrared diodes on the material's surface pixel distances become too large. On light adsorbent materials the depth values are very noisy.

The acoustic sensors included both a commercial (Acoustic Magic Voice Tracker Array™ Microphone) as well as a custom-made microphone array which were used simultaneously. The disadvantages of the commercial solutions are the rigid, firm arrangement, the fixed number and the position of the microphones. The enclosed software/driver regarding to the signal pre-processing is usually available only as a "Black Box" solution and there is only one output channel. This is unsuitable for testing, integrating and adapting our own algorithms for coming up with a high quality microphone arrangement. Therefore, it became necessary to use a custom-made microphone array that allows for flexibility in the algorithm development stage and optimal signal pre-processing. The custom-made array had eight equally spaced microphones. In all recording sessions we used both microphone arrays in parallel.

4. DATABASE IMPLEMENTATION

4.1. Equipment Setup

For the purpose of data collection two indoors and one outdoors recording sites were established. These sites cover the basic requirements of the target scenarios (smart-home living room, indoors public area, airport, and outdoor scenarios – ATM, airport, etc).

The indoors data collection sites were established on the premises of the University of Patras. Fig. 6 shows the smart-home living room from several sensor views.

During the two data collection campaigns (in June and July) we relied on a common strategy that utilized a distributed processing and storage architecture, with individual processing unit (IPU) and data repository for each sensor. The IPUs, which in our case were either laptop or desktop PCs with some additional hardware, served for management of the sensors and for provisional data storage. The use of IPUs was motivated by the dissimilar requirements of the specialized concomitant software involved in the sensor management and data collection. Another major factor was guaranteeing the data throughput requirements with some reserve, in order to prevent corruption or loss of incoming data streams.

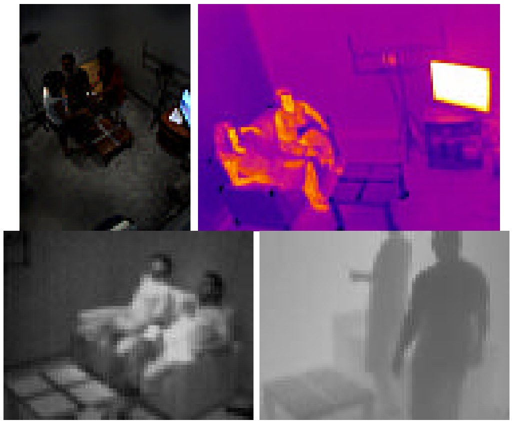

Fig. 6. Captured images of different narrow aisles, one entrance area with a larger door and one open area. Sensors used include video (top 5 images), thermal infrared (lower left) and a PMD device (lower right).

The heterogeneous nature of sensors and the different sampling rates required special care for proper time-stamping and data synchronization. Since synchronization among all sensors was a mandatory requirement, a number of solutions were studied. Short synchronization, which is recognizable in all sensors at a time, has been used as trigger. This event has been conducted by a person standing in sight of all visual sensors and performing a clap with his hands [5]. The clap could also be picked up by the utilized microphone arrays. In a subsequent processing step the data has been manually aligned. Applying multiple synchronization events varying frame rates could be aligned, by comparing the sampling rate between two or more detected events.

Due to the range of environments that the PROMETHEUS database had to cover, during different recording sessions we had to implement different equipment deployment schemes. These schemes differed in the sensor set employed in each setup, mainly due to the dissimilar application requirements but also due to the specifics of the indoor and outdoor environmental conditions. The equipment deployment setups for the different sessions were restricted to four basic schemes: two indoors and two outdoors. Thus, each setup followed one of the four predesigned schemes but intentionally varied slightly among the sessions, for better coverage of all aspects of the application scenarios. The indoor and outdoor scenario shared the same equipment and sensors but differed in the number of sensors, their particular deployment and the setup of their field of perception. Exceptions were the infrared motion detection sensor and the motion capture equipment which were used only in the indoor scenarios.

The most crucial part in sensor placement is the provision of an adequate perception field. This is usually applied to each scenario and there are few guidelines to follow. In wide open spaces with the tendency of overcrowding it is reasonable to install multiple cameras with differing field of view. This way, occlusions can be resolved easily.

All these sessions have been recorded with a total of nine video cameras, one stereo camera, two thermal sensors, one PMD and one microphone array with eight sensors. Five video cameras were used for overviews in similar positions as described above, as the same outdoor location has been used. Additional video cameras were placed on ground level to provide detail views of relevant locations. One thermal sensor was located in parallel to the PMD sensor and a video camera inside the building, for the initialization process, while an additional infrared sensor was placed on top of the building to enhance tracking in the global domain.

4.2. Annotation

In order to evaluate implemented algorithms and provide a common basis for comparison of concurring approaches ground truth data is inevitable. Hence the creation is rather time consuming the required data has to be defined carefully. In the first step the events to detect are annotated, as these are the most relevant ones for the tracking task. This task is performed with ANVIL [6], a tool for the annotation of multimodal dialogue. For each scenario a different annotation scheme in XML format is created. These contain the relevant information for each activity intended to pick up, such as: (i) approaching ATM, (ii) using ATM, (iii) approaching person at ATM, (iv) robbing person at ATM, (v) running away from ATM. Start and end of each activity was set with frame accuracy, by browsing a video or audio file. All events relevant to an individual are stored in a XML file.

For the needs of the tracking tasks, annotation on a frame level was required. In each frame the detected persons are compared to the manually labeled ground truth data. Though semi automated systems based on state of the art tracking systems [7] show reasonable high performance, errors have to be removed manually, which again requires the analysis of the data streams. The difficulty in this is to assign a unique ID to each individual, resulting in a trajectory. This constraint complicates the annotation process, and simple landmarking is not possible. Therefore specialized tools, like VIPER [8], have been used. It provides the possibility to label objects with polygons, points and bounding boxes which are assigned to a unique ID. Unfortunately VIPER is only designed for one singular view, which is rather inconvenient if persons are occluded by objects in one sight or leaving a field of view and reentering in another one. The British home office is releasing a multiple camera tracking scenario [9] with a VIPER based annotation tool limited to five cameras. Hence some of the recorded scenarios contained more visual sensors; a novel annotation tool has been implemented in MATLAB. It provides a graphical user interface to assign trajectories to individual and choose the adequate field of view. In order to save time and money only one view is required. The lowest and highest point of each object are landmarked within the frames and are subsequently transformed into world coordinates applying the planar homography constraint [10], as all sensors are calibrated. This way, trajectories can be created in the ground plane and transformed back into any other view, to evaluate 2D tracking.

Acoustic modality can play either a stand-alone or a complementary role towards detection and categorization of many types of human activities. A great quantity of sound events were captured during the PROMETHEUS recording campaign and afterwards manually annotated. The audio corpus is composed of (i) typical and atypical vocal reactions, (ii) typical and atypical non-vocalic sound events and (iii) background environmental noise. The particular sounds are encountered both in normal and abnormal situations and they are to be used for constructing representative probabilistic models to describe each sound category, which will be integrated in the final system. It should be noted that a similar audio corpus incorporating vocal and non-vocal sounds indicative of emergency situations does not exist while the one closer to our work is reported in [11] where the creation of a fiction database for emotion detection in abnormal situations was explained.

The tool that was used to annotate the PROMETHEUS multi-channel audio recordings was PRAAT [12]. The main advantages of the specific tool are that it offers multi-tier labeling. The annotation results are saved in TextGrid format, an easily-readable text file. Each recording is accompanied by one TextGrid file which contains the respective multi-tier annotation including the sound events and/or background noise that exist in the recordings. Both microphone arrays recorded signals in WAV format at the sampling rate of 32 kHz with 16 bit quantization.

The following tiers and tags were included during the audio annotation phase: (i) *Atypical Sound Event*, which referred to abnormal non-vocalic sounds with tags such as dropping of objects, fracture of material, footsteps, door sounds, fire alarm and other dominant events, (ii) *Typical Sound Event*, which corresponded to normal sounds such as door bell, normal speech, interaction with Socrates (smart room) etc, (iii) *Background Noise*, which was edited when a background noise appears such as wind, speech in the background, music or other noise, (iv) *Atypical Vocal Reactions*, which included human vocal sounds related to negative emotions while the tags were pain, fear, sorrow and anger, (v)

Sex, which was used only in one person typical and atypical speech (male/female), (vi) *Verbal*, used when speech audio events occur (yes/no) and (vii) *Audio Quality*, which reflected upon the quality of the audio signal with the next tags: clean, noise, music and other noise.

A comprehensive but simple and brief tutorial (in text and video format) was provided to five different annotators who worked on the particular task. It should be noted that the corresponding video files were given to them which were used whenever a clear decision upon the ongoing situation could not be made. The picture/video capability facilitates for proper acoustic event annotation and reduces the disagreement between annotators. As it was observed when video is available (in addition to the audio) the annotators tend to reach consensus on nearly all controversial fragments. In contrast to the audio only case, when a fragment is heard out of the context, humans tended to disagree in approximately 15% of the segments.

4.3. Database Organization

All the data that were gathered during the PROMETHEUS recording campaign has been categorized per scenario. The scenarios have been recorded more than one time with numerous variations and for each time the synchronized data sequences have been grouped together. An additional resource to the video files is their corresponding series of frames which are kept in a separate folder for serving annotation needs as well as further processing. Manually labeled annotation files in text (regarding audio files) and xml (regarding video files) format are also available with respect to each data file of different modality. Furthermore the data were split into test and train sets, with nine to one ratio in a random way.

5. CONCLUSIONS

This paper reported on the development of a new heterogeneous database that aims at supporting the research and development activities related to human behavior tracking and interpretation. The database was recorded in two indoors (smart-home, public area) and two outdoors (ATM, airport) setups. Each setup involved multiple video, thermal, and audio sensors, whose number and configuration was designed according to the scenarios of interest. The motivation behind the database design, as well as the implementation of the recording campaigns, the sensors used, the actual setups and the data annotation tools and procedures were described in details.

ACKNOWLEDGEMENTS

The work reported in this paper was done within contract FP7-214901 “Prediction and Interpretation of human behaviour based on probabilistic models and heterogeneous sensors”, PROMETHEUS project, co-funded by the European Commission.

The authors would like to acknowledge the contribution of Anna Linderhed, Fredrik Hemström, Tadeusz Olma, Paulo Menezes, Hadi Aliakbarpour, et al who had a major contribution in the implementation of the data collection campaigns. Finally, the authors would like to acknowledge the devotion of the professional actors implementing the action scenarios, and the effort of the team of annotators who spend hundreds of hours annotating the audio and video streams.

6. REFERENCES

[1] J. Ferryman, D. Tweed, “An Overview of the PETS 2007 Dataset,” PETS 2007, IEEE, 2007.

[2] D. Arsić, M. Hofmann, B. Schuller, G. Rigoll, “Multi-Camera Person Tracking and Left Luggage Detection Applying Homographic Transformation,” PETS 2007, IEEE, 2007.

[3] D. Thirde, L. Li, J. Ferryman, “An Overview of the PETS 2006 Dataset,” PETS’ 2005, pp. 317-324, 2005.

[4] F. Wallhoff, M. Ruß, G. Rigoll, J. Göbel, H., “Diehl Improved Image Segmentation Using Photonic Mixer Devices,” ICIP 2007, vol. 6, pp. 53–56, 2007.

[5] D. Lo, R.A. Goubran, R.M. Dansereau, “Multimodal talker localization in video conferencing environments,” HAVE 2004, pp. 195-200, 2004.

[6] M. Kipp, “Anvil - A Generic Annotation Tool for Multimodal Dialogue,” Eurospeech '01, pp. 1367-1370, 2001.

[7] D. Arsic, B. Schuller, G. Rigoll, “Multiple Camera Person Tracking in Multiple Layers Combining 2D and 3D Information,” M2SFA2’ 2008, Marseille, France, 2008.

[8] V.Y. Mariano, J. Min, J.-H. Park, R. Kasturi, D. Mihalcik, D. Doermann, T. Drayer, “Performance Evaluation of Object Detection Algorithms,” ICPR’2002, pp. 965-969, 2002.

[9] UK Home Office, “Multiple-camera tracking scenario (MCTS)”, October 2008. Available: http://scienceandresearch.homeoffice.gov.uk/hosdb/publications/cctv-publications/ MCTS_Scenario_Definition_Ma1.pdf?view=Binary

[10] S.M. Khan, M. Shah, “A Multiview Approach to Tracking People in Crowded Scenes Using a Planar Homography Constraint,” ECCV 2006, Graz, Austria, pp. 133-146, 2006.

[11] C. Clavel, I. Vasilescu, L. Devillers, T. Ehrette, “Fiction database for emotion detection in abnormal situations,” ICSLP’ 2004, Jeju, Korea, 2004.

[12] PRAAT software. Available: http://www.praat.org