AUDIOVISUAL BEHAVIOR MODELING BY COMBINED FEATURE SPACES

Björn Schuller, Dejan Arsic, Gerhard Rigoll

Technische Universität München
Inst. for Human-Machine Communication
Arcisstr. 21, 803333 München, Germany
schuller@IEEE.org

Matthias Wimmer, Bernd Radig

Technische Universität München
Department of Informatics
Boltzmannstr. 3, 85748 Garching, Germany
wimmerm@cs.tum.edu

ABSTRACT

Great interest is recently shown in behavior modeling, especially in public surveillance tasks. In general it is agreed upon the benefits of use of several input cues as audio and video. Yet, synchronization and fusion of these information sources remains the main challenge. We therefore show results for a feature space combination, which allows for overall feature space optimization. Audio and video features are thereby firstly derived as Low-Level-Descriptors. Synchronization and feature combination is achieved by multivariate time-series analysis. Test-runs on a database of aggressive, cheerful, intoxicated, nervous, neutral, and tired behavior in an airplane situation show a significant improvement over each single modality.

Index Terms— Audiovisual Emotion Recognition, Affective Computing, Synergistic Multimodality, Feature Fusion

1. INTRODUCTION

The related research fields of behavior and emotion recognition have recently grown an important factor in human-machine interfaces. Apart from many exciting applications especially the field of surveillance has lately gained more interest. It seems commonly agreed that a fusion of several input cues is advantageous [1], yet most efforts are spent on uni-modal approaches [2]. The main problem remains synchronization and synergistic fusion of the streams. This comes, as speech is mostly processed at turn-level while vision based emotion or behavior modeling mostly operates at a constant frame or macro-frame-basis. However, we recently demonstrated that the analysis of speech at such a constant rate is less reliable [3]. On the other hand vision results are mostly synchronized by e.g. majority voting to map frame results on a turn-level. Likewise most works unite audio and video in a late semantic fusion. Yet, many advantages of an early feature fusion are known as keeping all knowledge for the final decision process and the ability of a combined feature-space optimization. We therefore suggest the use of multivariate time-series-analysis as typically used in speech emotion recognition for combined audio- and video-processing to realize a combined pass of both low-level descriptors (LLD) resulting in such an early fusion. The paper is structured as follows: sec. 2 and sec. 3 introduce our LLD for audio and video. Next, a combined multivariate time-series analysis by means of descriptive statistical analysis is described in sec. 4 and the optimization of the feature space is discussed in sec. 5. Finally, we introduce a novel audiovisual database for airplane behavior modeling in sec. 6 prior to results in sec. 7 and discussion in sec. 8.

2. AUDIO LOW-LEVEL-DESCRIPTORS

In former works [4] we compared static and dynamic feature sets for the prosodic analysis and demonstrated the higher performance of derived static features by multivariate time-series analysis. As an optimal set of such global features is broadly discussed [1, 2], we consider an initially large set of 38 LLD which cannot all be described in detail here. However, the target is to become utmost independent of the spoken content and ideally also of the speaker, but model the underlying emotion with respect to prosodic, articulatory and voice quality aspects. The feature basis is formed by the raw contours of zero crossing rate (ZCR), pitch, first seven formants, energy, spectral development, and Harmonics-to-Noise-Ratio (HNR). Duration based features rely on common bi-state dynamic energy threshold segmentation and voicing probability. In order to calculate the according LLD 20 ms frames of the speech signal are analyzed every 10 ms using a Hamming window function. Pitch is detected by the auto correlation function (ACF) with window compensation and dynamic programming (DP) for global error minimization. HNR also relies on the ACF. The values of energy resemble the logarithmic mean energy within a frame. Formants base on 18-point LPC spectrum and DP. We use their position and bandwidth, herein. For spectral development we use 15 MFCC coefficients and an FFT-spectrum out of which we calculate spectral flux, centroid and 95%-roll-off-point after dB(A)-correction according to human perception. Low-pass SMA filtering smoothes the raw contours prior to the statistical analysis. First and second order regression coefficients are subsequently calculated for selected LLDs resulting in a total of 88 audio LLD.

3. VIDEO LOW-LEVEL-DESCRIPTORS

Model-based image interpretation techniques serve as a good workhorse for facial expression recognition, see Figure 3. Those methods exploit a priori knowledge of objects, for instance texture or shape. This knowledge is described by a small number of model parameters. Reducing the large amount of information in images to a small set of parameters facilitates and accelerates further image interpretation, such as facial expression recognition. Model fitting is the computational challenge of finding the model parameter values whose corresponding model configuration describes the content of the image best [5]. Model-based image understanding methods consist of four main components: the model, the fitting algorithm, the objective function, and the interpretation phase.

The *model* consists of a parameter vector $\mathbf{p}$ that represents the possible configurations of the model, such as position, rotation, scal-

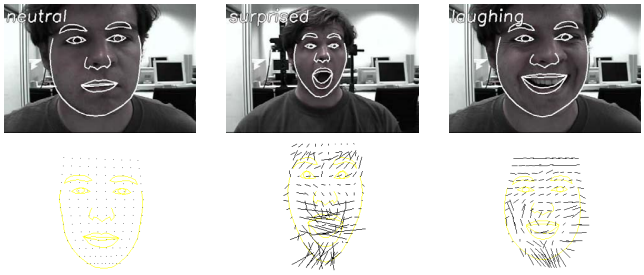

Fig. 1. Fitting a deformable face model to images and inferring different facial expressions by taking structural and temporal image features into account.

ing, and deformation. Furthermore, models define their mapping onto the 2D surface of an image, such as to a set of feature points, a contour, or a textured region. Our approach integrates a deformable model, which has been introduced by Cootes et al. [6] with the name *Point Distribution Model*. Referring to [7], deformable models are very suitable for analyzing the variations that occur within a human face. The model's parameter vector $\mathbf{p} = (\delta_x, \delta_y, s, \alpha)^T$ consists of the translation δ_x and δ_y, the scaling factor s, the rotation α, and a vector of deformation parameters $\mathbf{b} = (b_1, ..., b_n)^T$. It infers the pose, the opening of the mouth, the roundness of the eyes, or the raising of the eye brows, as depicted in Figure 3. In this work, we set $n = 17$ in order to cover the necessary modes of variation with the face model. For a detailed explanation of this type of model, we refer to [6].

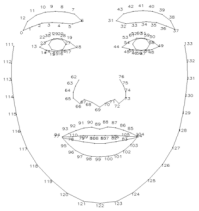

Fig. 2. Face model

The *objective function* evaluates how good the fit of a model parametrization to an image is. Examples and synonyms of the objective function are likelihood function, similarity function, energy function, cost function, goodness function, and quality function. Depending on its definition that function needs either be minimized or maximized. Traditionally, the calculation rules of objective functions are obtained by selecting a small number of simple image features, such as edges or corners, which are mathematically combined by human intuition. Wimmer et al. [8] show that this methodology is erroneous and tedious and propose a novel approach, which learns the objective function from annotated example images.

The *fitting algorithm* searches for those model parameters that describe the visible face in the image best. Therefore, the fitting algorithms needs to find that parameter vector $\mathbf{p}$ that minimizes the objective function. Since our methods are independent of the fitting algorithm, we shall not elaborate on them in this paper but refer to [5] for a recent overview and categorization of fitting algorithms.

The *interpretation phase* extracts structural features from the image and the model and temporal features from the image sequence from which the intended information is derived, which are the visible

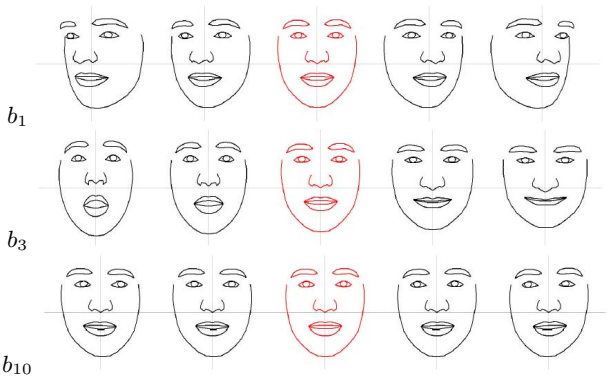

Fig. 3. This figure illustrates the huge extent of deformation that are applied to the model by changing just one parameter in each row. Each parameter is manipulated within a certain extent by keeping the others fixed. Topmost row: manipulating parameter b_1 changes the rotation of the head. Middle row: Parameter b_3 opens the mouth. Lowermost row: Parameter b_{10} moves the pupils. Note that the pupils are both moved in accordance to each other.

facial expressions in our case. A comprehensive overview is found in [9, 10].

Fig. 4. Deriving skin color from the camera image. Middle: non-adaptive. Right: adaptive to the person and to the context.

A system for recognizing facial expression that runs autonomously needs to detect faces within the images without human interference. We detect the presence of a face using classifiers that take Haar-like image features, which was proposed in [11]. We extend this approach and train it to recognize several facial parts, too, such as the eyes, the nose and the mouth. As the result of a subsequent application of those classifiers we know the rough model parameters, i.e. the rough location of the visible face in the image. Applying a fitting algorithm thereafter greatly refines the estimation of the model parameters. The refinement is essential for our system, because our system relies on exact knowledge of the constitution of the face and its facial parts. Fitting algorithms have been the subject of intensive research and evaluation. Complex algorithms are the result of that vast research, e.g. *Simulated Annealing*, *Genetic Algorithms*, *Particle Filtering*, *RANSAC*, *CONDENSATION*, and *CCD*. But they still do not suffice the need of model-based image understanding applications by far. In contrast, the objective function is usually determined ad hoc and heuristically, using the designer's intuitions about a good measure of fitness. Afterwards, its appropriateness is subjectively determined by inspecting the result of the objective function on example images and example parametrizations of the model. If the result is not satisfactory the function is tuned or redesigned from scratch. In short, the traditional way of designing objective functions is rather an art than a science. To avoid the problems of the iterative design approach, we stick to the approach of Wimmer et al. [8] and learn the objective function from annotated training images. This splits up the generation of the objective function into several inde-

pendent pieces, which are partly automated. This novel methodology has several benefits: First, automated steps mainly replace the labor-intensive design of the objective function. Second, this approach is less error-prone, because giving examples of good fit is much easier than explicitly specifying rules that need to cover all examples. Third, this approach does not need any expert knowledge and therefore it is generally applicable and not domain-dependent. The bottom line is that this yields more robust and accurate objective functions, which greatly facilitate the task of the fitting algorithms. Skin color is an important feature of faces and our system for recognizing facial expressions vastly benefits from robust skin color detection. But skin color may look quite differently, depending on camera settings, illumination, shadows, people's tans, ethnic groups. That variation is a challenging aspect of skin color classification. Wimmer et al. [12] present an approach that uses a high level vision module in order to detect an image specific skin color model. This color model is representative for the context conditions within the image and is used to adapt dynamic skin color classifiers to it. This approach facilitates to distinguish skin color from very similar color like lip color or eyebrow color, see Figure 4. Since the borders of skin color regions are well determined and the contour of our face model represents those borders, this approach greatly supports the process of model fitting. The mentioned technology for learning the objective function and adaptively extracting skin color allows our approach to integrate a simple and quick fitting algorithm that is capable for real-time execution. It works like a *greedy algorithm* by searching the best matching model parameters from a local point of view. In other words, we search the minimum of the objective function via a hill-climbing strategy. Facial expressions are generally characterized from two aspects: On the one hand, they turn the face into a distinctive state, see upper row of Figure 3, and on the other hand, the hereby involved muscles show a distinctive motion, see lower row of Figure 3. In this work, we make use of both aspects of interpretation in order to obtain a profound basis of data to decide on. This section describes the computational approach of gathering information related to each aspect. We will obtain structural features in order to describe the first aspect, and temporal features to allow for the second aspect. Referring to the structural features, the constitution of the face is determined by the deformation parameters $\mathbf{b}$. The values of the model parameters δ_x, δ_y, s, and α have no influence on the person's facial expression and therefore they are not selected by this phase. Referring to the temporal features, we focus on the muscular activity of facial expressions, which is determined by the optical flow, see [13]. Since real-time issues are an important factor to surveillance tasks, the optical flow is just evaluated at a small number of locations within the face. This is achieved by deploying a 10×14 grid, which is tightly connected to the face model, w.r.t. its position, its rotation, and its size. During a short time period each grid point sums up the visible motion $g_{x,i}$ and $g_{y,i}$ with $1 \leq i \leq 140$. This period has been set to 2s in order to cover even slowly expressed emotions. The resulting LLD feature vector time series $\mathbf{t}$ that describes the current conditions is assembled from the structural features and the temporal features mentioned above. The structural features describe the model that fits to the current image and the temporal features describe the motion of the image sequence during the last 2 seconds.

$$\mathbf{t} = (b_1, ..., b_{10}, g_{x,1}, g_{y,1}, ..., g_{x,140}, g_{y,140})^T \quad (1)$$

4. MULTIVARIATE TIME-SERIES ANALYSIS

So far we extracted LLD as base-contours considering audio and video information on a raw level. As stated in sec. 2 these can be directly processed by dynamic modeling as Hidden Markov Models (HMM) or Dynamic Bayesian Nets (DBN). Yet, streams usually need to be synchronized for this purpose. We therefore prefer the application of functionals f to the LLD F in order to obtain a static feature vector x:

$$f : F \rightarrow \Re^1 \quad (2)$$

The higher level features are likewise derived by means of descriptive statistical analysis as linear moments, extremes, ranges, quartiles, or durations, and normalized. Overall the final per-turn feature vector consists of 276 audio (see tab. 1), and 1,048 video features.

Type	Pitch	Energy	Duration	Formant
[#]	12	11	5	105
Type	HNR	MFCC	FFT	ZCR
[#]	3	120	17	3

Table 1. Distribution of audio features

This feature vector x is now classified by use of Support Vector Machines (SVM) with polynomial Kernel and a couple-wise multi-class discrimination strategy. We decided for SVM as these proved the optimal choice in our extensive classifier comparison in speech-based emotion recognition [14].

5. FEATURE SPACE OPTIMIZATION

Apart from the choice of an optimal classifier also selection of the most relevant features is important as it saves computation time considering real-time processing and boosts performance as some classifiers are susceptible to high dimensionality. Therefore search for the right features seems mandatory. We chose Sequential Forward Floating Search (SFFS) with SVM as wrapper - that is employing the classifier's error as optimization criterion - within audio and video feature selection as it proved a reasonable compromise compared to NP-hard exhaustive search and proved a good choice in former works [14]. The search is performed by hill-climbing with forward and backward steps eliminating and adding features in a floating manner to an initially empty set. By the employed wrapper search a set is optimized as a whole rather than finding single attributes of high relevance. As an audiovisual super vector is constructed, we can select features also in one pass. The ranking of this selection directly points at the importance of audio and video features, each. The optimal number of features is afterward determined in accordance to the highest observed accuracy throughout selection.

6. AIRPLANE BEHAVIOR CORPUS

As public audiovisual emotion or behavior data is sparse, we decided to record a database crafted for the special target application of public transport surveillance. In order to obtain data in equivalent conditions of several subjects of diverse classes we decided for acted behavior. There is a broad discussion in the community with respect to acted vs. spontaneous data, which we will not address herein. However, it is believed, that mood induction procedures favor realism in behavior. Therefore a script was used, which leads the subjects through a guided storyline: prerecorded announcements by five different speakers were automatically played back controlled by a hidden test-conductor. As a general framework a vacation flight with return flight was chosen, consisting of 13 and 10

scenes as start, serving of wrong food, turbulences, falling asleep, conversation with a neighbor, or touch-down. The general setup consisted of an airplane seat for the subject positioned in front of a blue screen. Camera and a condenser microphone AEG 1000S MK II were fixed without occlusions of the subject. 8 subjects in gender-balance from 25a to 48a (mean 32a) took part in the recording. The language throughout recording is German. A total of 11.5h video was recorded and annotated independently after pre-segmentation by three experienced male labelers within a closed set as seen in tab. 3. This table also shows the final distribution of samples with total inter-labeler-agreement. This set will be referenced as ABC (Airplane Behavior Corpus) in the ongoing. The average length of the 396 clips in total is 8.4s.

7. EXPERIMENTS

In the research of behavior or emotion recognition data is usually sparse. As most popular evaluation strategy j-fold stratified cross validation (SCV) can therefore be named: SCV allows for testing and disjunctive training on the whole corpus available. We therefore use 10-fold SCV in the ongoing and present mean accuracies throughout cross-folds.

Tab. 2 shows our results for each single information source and the fusion of these. Features are firstly selected by SVM-SFFS as described in sec. 5, separately for audio and video as a pre-selection step to keep computation effort within reasonable limits. Subsequently, the combined set is reduced by another SVM-SFFS selection. Thereby the 250 audiovisual features are reduced to 55% set-size. As can be seen in the table, audio standalone is superior to video standalone. However, a remarkable overall gain is observed for the fusion of these two sources.

	Audio	Video	Audiovisual
Features [#]	93	157	139
Accuracy [%]	73.7	61.1	**81.3**

Table 2. Accuracies for modalities and fusion using SVM in a 10-fold SCV, database ABC.

Tab. 3 shows confusions of the final test-run with the optimized audiovisual feature set. As can be seen, most confusions occur between nervous and neutral, and intoxicated and cheerful behavior. Intoxicated behavior thereby is a complex behavior, as it can be aggressive as well as joyful. In the table also f_1-measures are presented to show the ratio between precision and recall for each class. Apart from aggressiveness, which is recognized most reliably as in most speech emotion recognition tasks, all classes are recognized with balanced reliability.

ground truth	aggressive	cheerful	intoxicated	nervous	neutral	tired	[#]	f_1 [%]
aggressive	83	1	0	1	2	0	87	91.7
cheerful	6	87	1	3	2	1	100	82.9
intoxicated	0	8	19	1	3	0	31	73.1
nervous	2	5	0	49	13	1	70	73.7
neutral	2	8	0	4	52	2	68	74.3
tired	1	1	1	5	0	32	40	84.2

Table 3. Behavior confusions and f_1-measures by use of SVM in a 10-fold SCV, optimized audiovisual feature set, database ABC.

8. DISCUSSION AND FUTURE WORK

The results presented in sec. 7 clearly show the superiority of a combined audiovisual approach. Interestingly, the total number of features could be further reduced by the combined feature selection. This also leads to overall higher accuracy, and the combined time-series-analysis approach to audiovisual behavior modeling proved highly promising.

In future work we aim at database-enlargement, tests on further data-sets, and in-depth feature analysis.

9. REFERENCES

[1] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Processing magazine*, vol. 18, no. 1, pp. 32–80, January 2001.

[2] M. Pantic and L. Rothkrantz, "Toward an affect-sensitive multimodal human-computer interaction," *Proccedings of the IEEE*, vol. 91, pp. 1370–1390, September 2003.

[3] B. Schuller and G. Rigoll, "Timing levels in segment-based speech emotion recognition," in *Proc. INTERSPEECH 2006, ICSLP*, Pittsburgh, USA, 2006, ISCA.

[4] B. Schuller, G. Rigoll, and M. Lang, "Hidden markov model-based speech emotion recognition," in *Proc. ICASSP 2003*, 2003, vol. II, pp. 1–4.

[5] R. Hanek and M. Beetz, "The contracting curve density algorithm: Fitting parametric curve models to images using local self-adapting separation criteria," *International Journal of Computer Vision (IJCV)*, vol. 59, no. 3, pp. 233–258, 2004.

[6] T. F. Cootes and C. J. Taylor, "Active shape models – smart snakes," in *Proc. of the* 3^{rd} *British Machine Vision Conference 1992*. 1992, pp. 266 – 275, Springer Verlag.

[7] G. J. Edwards, T. F. Cootes, and C. J. Taylor, "Face recognition using active appearance models," vol. LNCS-Series 1406–1607, pp. 581–595, 1998.

[8] M. Wimmer, F. Stulp, S. J. Tschechne, and B. Radig, "Learning robust objective functions for model fitting in image understanding applications," in *British Machine Vision Conference 2006*, Edinburgh, Great Britain, Septembre 2006, vol. 3, pp. 1159 – 1168, BMVA.

[9] M. Pantic and L. Rothkrantz, "Automatic analysis of facial expressions: The state of the art," *IEEE Transactions on PAMI*, vol. 22, no. 12, pp. 1424–1445, 2000.

[10] I. Cohen, N. Sebe, L. Chen, A. Garg, and T. Huang, "Facial expression recognition from video sequences: Temporal and static modeling," 2003.

[11] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *CVPR*, 2001.

[12] M. Wimmer, B. Radig, and M. Beetz, "A person and context specific approach for skin color classification," in *Proc. ICPR 2006*, http://www.ieee.org/, August 2006, vol. 2, IEEE.

[13] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proc. IJCAI '81*, April 1981, pp. 674–679.

[14] B. Schuller, R. Mller, M. Lang, and G. Rigoll, "Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensembles," in *Proc. Interspeech 2005*, Lisboa, Portugal, 2005, ISCA.