LOW-LEVEL FUSION OF AUDIO AND VIDEO FEATURE FOR MULTI-MODAL EMOTION RECOGNITION

Matthias Wimmer
Perceptual Computing Lab, Faculty of Science and Engineering, Waseda University, Tokyo, Japan
matthias.wimmer@cs.tum.edu

Björn Schuller, Dejan Arsic, Gerhard Rigoll
Institute for Human-Machine Communication, Technische Universität München, Germany

Bernd Radig
Chair for Image Understanding, Technische Universität München, Germany

Keywords: Emotion Recognition, Audio-visual Processing, Multi-modal Fusion

Abstract: Bimodal emotion recognition through audiovisual feature fusion has been shown superior over each individual modality in the past. Still, synchronization of the two streams is a challenge, as many vision approaches work on a frame basis opposing audio turn- or chunk-basis. Therefore, late fusion schemes such as simple logic or voting strategies are commonly used for the overall estimation of underlying affect. However, early fusion is known to be more effective in many other multimodal recognition tasks.
We therefore suggest a combined analysis by descriptive statistics of audio and video Low-Level-Descriptors for subsequent static SVM Classification. This strategy also allows for a combined feature-space optimization which will be discussed herein. The high effectiveness of this approach is shown on a database of 11.5h containing six emotional situations in an airplane scenario.

1 INTRODUCTION

Automatic recognition of human emotion has recently grown an important factor in multimodal human-machine interfaces and further applications. It seems commonly agreed that a fusion of several input cues is advantageous, yet most efforts are spent on uni-modal approaches (Pantic and Rothkrantz, 2003). 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. In speech processing, a turn denotes an entire phrase or a similar contiguous part of the audio stream. (Schuller and Rigoll, 2006) shows that the analysis of speech at such a constant rate is less reliable. For this reason, vision and audio results are mostly synchronized by late fusion, e.g. majority voting, to map frame results to a turn-level-based interpretation. Likewise, most works unite audio and video in a late semantic fusion.

As addressed in this paper, early feature fusion is known to provide many advantages, such as keeping all knowledge for the final decision process and the ability of a combined feature-space optimization. We therefore suggest to statistically analyzing multivariate time-series as used in speech emotion recognition for a combined processing of video-based and audio-based low-level descriptors (LLDs). This approach represents early feature fusion, which promises to exploit more semantical information from the given data and thus provides more accurate results.

The paper is structured as follows: Section 2 and Section 3 explain the acquisition of LLDs for video and audio. Section 4 describes the functional-based analysis and the optimization of the combined feature space. Section 5 introduces the evaluation data and elaborates on our experiments conducted. Summary and outlook is finally given in Section 6.

Figure 1: Model-based techniques greatly support the task of facial expression interpretation. The parameters of a deformable model give evidence about the currently visible state of the face.

From: Proc. VISAPP 2008, Funchal, Madeira, Portugal
INSTICC, ISBN 978-989-8111-21-0

2 VIDEO LOW-LEVEL DESCRIPTORS

Model-based image interpretation exploits a priori knowledge about objects, such as the shape or the texture of a human face. Therefore, these techniques serve as a good workhorse for extracting the vision-based LLDs from our emotion recognition system, see Figure 1. They reduce the large amount of image data to a small set of model parameters, which facilitates and accelerates image interpretation. Model fitting is the computational challenge of finding the model parameterization that best describe a given image. Our system consists of six components that are common parts of model-based image interpretation: the model itself, the localization algorithm, the skin color extraction, the objective function, the fitting algorithm, and the extraction of the video-based LLDs.

The model contains a parameter vector $\mathbf{p}$ that represents the possible configurations of the model, such as position, orientation, scaling, and deformation. They are mapped onto the surface of an image via a set of feature points, a contour, a textured region, etc. Referring to (Edwards et al., 1998), deformable models are highly suitable for analyzing human faces with all their individual variations. Our approach makes use of a statistics-based deformable model, as introduced by (Cootes and Taylor, 1992). The model parameters $\mathbf{p} = (t_x, t_y, s, \theta, \mathbf{b})^T$ contain the translation, the scaling factor, the rotation, and a vector of deformation parameters $\mathbf{b} = (b_1, ..., b_m)^T$. The latter component describes the facial pose, opening of the mouth, roundness of the eyes, raising of the eye brows, etc., see Figure 3. In this work, we set $m = 17$ in order to cover all necessary modes of variation.

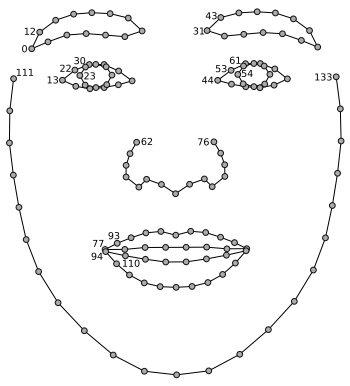

Figure 2: Our deformable model of a human face consists of 134 contour points and represents the major facial components.

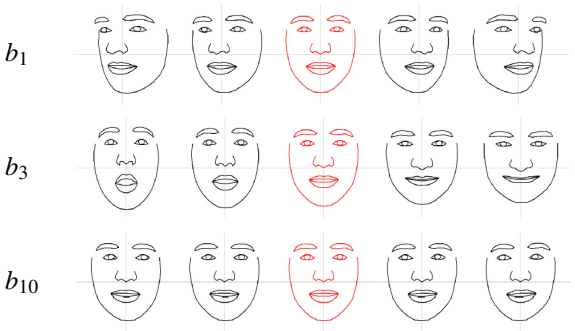

Figure 3: Changing individual model parameters yields highly semantic facial deformation. Top-most row: b_1 affects the orientation of the head. Center row: b_3 opens the mouth. Lower-most row: b_{10} moves pupils accordingly.

The localization algorithm automatically starts the interpretation process in case the intended object is visible. It computes an initial estimate of the model parameters that is further refined by the subsequent fitting algorithm. Our system integrates the approach of (Viola and Jones, 2004), which is able to detect the affine transformation parameters (t_x, t_y, s, and θ) of our 2D face model in case the image shows a frontal view face.

We also integrated the ability to roughly estimate the deformation parameters $\mathbf{b}$ to obtain higher accuracy. For this reason, we apply a second iteration of the Viola and Jones object detector to the previously determined image region that contains the face. We specifically learned the algorithm to localize the facial parts, such as eyes and mouth, within this iteration. In the case of the eyes, our positive training images show the eye region only, whereas the negative training images contain the vicinity of the eyes, such as the cheek, the nose, or the brows. Note that the resulting eye detector[1] is not able to extract the eyes from a complex image, because most content of these images was not part of the training data. However, the it is highly appropriate to determine the location of the eyes given a pure face image or a face region within a complex image. To some extent, this approach is similar to the Pictoral Structures, Felzenszwalb et al. (Felzenszwalb and Huttenlocher, 2000) elaborate on, because we also define a tree-like structure where a superordinate element (face) contains the subordinate elements (eyes, mouth, etc.) and where a geometric relation between these elements is given.

Skin color extraction acquires reliable information

[1] Our system integrates object detectors for several facial parts. We make them accessible at the following web page: www9.in.tum.de/people/wimmerm/se/project.eyefinder

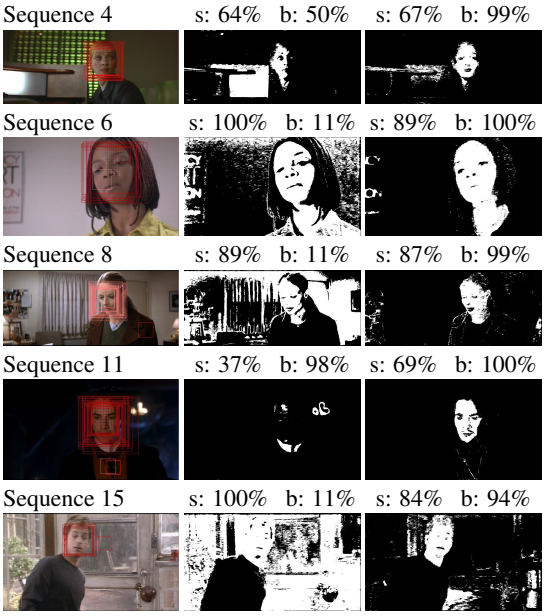

Figure 4: Deriving skin color from the camera image (left) using the non-adaptive classifier (center) and adapting the classifier to the person and to the context (right). The numbers indicate the percentage of correctly identifying skin color (s) and the background (b). These images have been extracted from some image sequences of the Boston University Skin Color Database (Sigal et al., 2000).

about the face and the facial components, as opposed to pixel values. It gives evidence about the location and the contour lines of skin colored parts, on which subsequent steps rely. Unfortunately, skin color varies with the scenery, the person, and the technical equipment, which challenges the automatic detection. As in our previous work (Wimmer et al., 2006), a high level vision module determines an image-specific skin color model, on which the actual process of skin color classification bases. This color model represents the context conditions of the image and dynamic skin color classifiers adapt to it. Therefore, our approach facilitates to distinguish skin color from very similar color, such as lip color or eyebrow color, see Figure 4. Our approach makes use of this concept, because it clearly extracts the borders the skin regions and subsequent steps fit the contour model to these borders with high accuracy.

The objective function $f(I, \mathbf{p})$ yields a comparable value that specifies how accurately a parameterized model $\mathbf{p}$ matches an image I. It is also known as the likelihood, similarity, energy, cost, goodness, or quality function. Without losing generality, we consider lower values to denote a better model fit. Traditionally, objective functions are manually specified by first selecting a small number of simple image features, such as edges or corners, and then formulating mathematical calculation rules. Afterwards, the appropriateness is subjectively determined by inspecting the result on example images and example model parameterizations. If the result is not satisfactory the function is tuned or redesigned from scratch. This heuristic approach relies on the designer's intuition about a good measure of fitness. Our earlier publications (Wimmer et al., 2007b; Wimmer et al., 2007a) show that this methodology is erroneous and tedious.

To avoid this, we propose to learn the objective function from annotated example images. Our approach splits up the generation of the objective function into several independent steps that are mostly automated. This provides several benefits: first, automated steps 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 approach yields more robust and accurate objective functions, which greatly facilitate the task of the fitting algorithms. For a detailed description of our approach, we refer to (Wimmer et al., 2007b)

The fitting algorithm searches for the model that best describes the face visible in the image. Therefore, it needs to find the model parameters that minimize the objective function. Fitting algorithms have been the subject of intensive research and evaluation, e.g. *Simulated Annealing*, *Genetic Algorithms*, *Particle Filtering*, *RANSAC*, *CONDENSATION*, and *CCD*. We refer to (Hanek, 2004) for a recent overview and categorization. Since we adapt the objective function rather than the fitting algorithm to the specifics of the face interpretation scenario, we are able to use any of these standard fitting algorithms.

Emotion interpretation applications mostly require real-time capabilities, our experiments in Section 5 have been conducted with a quick *hill climbing* algorithm. Note that the reasonable specification of the objective function makes this local optimization method nearly as accurate as global optimization strategies.

The extraction of vision LLDs infers information that is descriptive for facial expressions considering the content of the current image and the entire image sequence as well as the model parameters. Two aspects characterize facial expressions: first, they turn the face into a distinctive state (Littlewort et al., 2002)

Figure 5: Model-based image interpretation for facial expression recognition: Fitting a deformable face model to images and inferring different facial expressions by taking structural and temporal image features into account.

and second, the involved muscles show a distinctive motion (Schweiger et al., 2004; Michel and Kaliouby, 2003). Our approach considers either aspect by extracting both structural and temporal features. This large amount of feature data provides a fundamental basis for the subsequent sensor fusion step and, in turn, for recognizing human emotion.

Structural features: The deformation parameters **b** describe the constitution of the visible face. The examples in Figure 3 illustrates the relation between the facial expression and the components of **b**. Since it provides structural information, we consider **b** for the interpretation process. In contrast, the affine transformation parameters t_x, t_y, s, and θ do not give evidence about the facial expression. They represent the position and orientation of the face model instead and therefore, we do not consider them as features for the interpretation process.

Temporal features: Facial expressions also emerge from muscle activity and therefore, the motion of particular feature points within the face is descriptive as well. Again, real-time capability is important and therefore, a moderate number of feature points within the area of the face model is considered only. The relative location of these points is connected to the structure of the face model. Note that we do not specify these locations manually, because this assumes a good experience of the designer in analyzing facial expressions. In contrast, we automatically generate a moderate number of G feature points that are equally distributed, see Figure 5. We expect these points to move uniquely and predictably in the case of a particular facial expression. As low-level features, we sum up the motion $g_{x,i}$ and $g_{y,i}$ of each point $1 \leq i \leq G$ during a short time period. We set this period to 2 seconds to cover slowly expressed emotions as well. The motion of the feature points is normalized by the affine transformation of the entire face (t_x, t_y, s, and θ) in order to separate the facial motion from the rigid head motion.

$$\mathbf{t}_v = (b_1, ..., b_{17}, g_{x,1}, g_{y,1}, ..., g_{x,140}, g_{y,140})^T \qquad (1)$$

The vision-based LLD feature vector $\mathbf{t}_v$ describes the currently visible face and it is assembled from the structural and the temporal features mentioned. The time series $\mathbf{T}_v$ is constructed from a sequence of $\mathbf{t}_v$ sampled at frame rate. It is established for a certain amount of time which is determined by speech processing.

3 AUDIO LOW-LEVEL DESCRIPTORS

In our former publication (Schuller et al., 2005), 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 by (Pantic and Rothkrantz, 2003), we consider an initially large set of 38 audio-based LLDs, 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 low-level descriptors, we analyze 20 ms frames of the speech signal 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 a 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 features.

These low-level descriptors are combined to the audio-based feature vector $\mathbf{t}_a$. Again, the time series $\mathbf{T}_a$ is constructed from sampling $\mathbf{t}_a$ over a certain amount of time.

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

Table 1: Distribution of acoustic features

4 EMOTION CLASSIFICATION

The preceding sections shows the extraction of raw audio and video low-level descriptors. Here, we describe the fusion of these features in our early fusion approach.

4.1 COMBINING AUDIO AND VIDEO DESCRIPTORS

As stated in Section 3, these LLDs 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 combined low-level descriptors $\mathbf{T}_c$=$[\mathbf{T}_v, \mathbf{T}_a]$ in order to obtain a feature vector $\mathbf{x} \in \mathbb{R}^d$, see Equation 2. As opposed to the time-series data, this feature vector is of constant dimension d, which allows for an analysis with standardized techniques.

$$f : \mathbf{T}_c \rightarrow \mathbf{x} \tag{2}$$

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 features, see Table 1, and 1,048 video features.

This feature vector $\mathbf{x}$ is now classified by use of Support Vector Machines (SVM) with polynomial Kernel and a couple-wise multi-class discrimination strategy.

4.2 OPTIMIZING FEATURE SPACE

Apart from the choice of an optimal classifier, selection of the most relevant features is important as well. It saves computation time considering real-time processing and boosts performance as some classifiers are susceptible to high dimensionality. We chose Sequential Forward Floating Search (SFFS) with SVM as wrapper to employ classification error as optimization criterion and avoid NP-hard exhaustive search recommended in (Schuller et al., 2005). A set is likewise optimized rather than finding single attributes of high relevance. As an audiovisual super vector is constructed, we can select features in one pass to point out the importance of audio and video features, each. The optimal number of features is determined in accordance to the highest accuracy throughout selection.

5 EXPERIMENTAL EVALUATION

This section describes the evaluation conducted upon the system introduced. Since there is no sufficient public data for our purpose, we acquired a sufficient data base by our own.

5.1 AIRPLANE BEHAVIOR CORPUS

As public audiovisual emotion data is sparse, we decided to record a database, which is crafted for our special target application of public transport surveillance. To obtain data in equivalent conditions of several subjects of diverse classes we decided for acted behavior, see Table 2. 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 by automatically played prerecorded announcements. The framework is a vacation flight with return flight, consisting of 13 and 10 scenes respectively, such as takeoff, serving of wrong food, turbulences, falling asleep, conversation with a neighbor, or touch-down. The setup is an airplane seat for the subject in front of a blue screen. A camera and a condenser microphone were fixed without occlusions of the subject.

In the acquizition phase, 8 subjects in gender-balance from 25 years to 48 years with a mean of 32 years took part. A total of 11.5 hours of video was recorded, pre-segmented, and annotated by 3 experienced labelers independently with a closed set as seen in Table 3. This segmentation process yields a total of 396 clips that contain both emotional audio and video data with an average length of 8.4 seconds.

	aggressive	cheerful	intoxicated	nervous	neutral	tired	TOTAL
[#]	87	100	31	70	68	40	**396**

Table 2: Distribution of behaviors, database ABC.

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.

This table also shows the final distribution with total inter-labeler-agreement. This set is referenced as ABC (Airplane Behavior Corpus).

5.2 EXPERIMENTS

We use j-fold stratified cross validation (SCV), because it allows for testing and disjunctive training on the whole corpus available. Table 3 shows individual class-wise f_1-measures for each feature stream and optimization with respect to combined and individual strategy. Most confusions occur between nervous and neutral, and intoxicated and cheerful behavior. Note that intoxicated behavior is a complex behavior, as it can be aggressive as well as joyful.

Table 4 summarizes the results: Features are firstly selected by SVM-SFFS as described in Section 4.2, separately for audio and video as a preselection step to keep computation effort in reasonable limits. Subsequently, the combined set is reduced by another SVM-SFFS selection. As can be seen audio standalone is superior to video standalone. However, a remarkable overall gain is observed for the fusion of these two sources. Table 3 illustrates the confusion of our classification scheme with respect to the different emotional states. These results clearly show the superiority of the combined audiovisual approach.

According to the description in Section 4.2, we further reduce the total number of features by the combined feature selection. Table 4 shows that this also leads to overall higher accuracy, and the combined time-series-analysis approach to audiovisual behavior modeling proved highly promising.

6 SUMMARY AND OUTLOOK

Former publications show both that early sensor fusion is advantageous over late fusion and that the integration of audio and video information greatly supports emotion recognition. However, previous approaches mostly apply late sensor fusion to this application because of the obstacles that these different types of sensor information pose.

The presented approach integrates state-of-the-art techniques in order to acquire a large range of both audio and video low-level features at frame rate. It applies well-known functionals to obtain representative and robust feature set for emotion classification. Our experiments show that this combined feature set is superior over the individual audio or video feature set. Furthermore, we conduct feature selection that again indicates that the combination outperforms the stand-alone approaches.

We are currently conducting explicit comparison to late fusion approaches that empirically prove our statements. In future work, we aim at testing our approach on further data sets and in-depth feature analysis. We will also investigate the accuracy and runtime performance of asynchronous feature fusion and the application of hierarchical functionals. Furthermore, we intend to dynamically model the emotional expression on a meta basis including both video and audio aspects.

ACKNOWLEDGEMENTS

This research is partly funded by a JSPS Postdoctoral Fellowship for North American and European Researchers (FY2007).

It has been jointly conducted by the Perceptual Computing Lab of Prof. Tetsunori Kobayashi at Waseda University, the Chair for Image Understanding at the Technische Universität München, and the Institute for Human-Machine Communication at the Technische Universität München.

REFERENCES

Cootes, T. F. and Taylor, C. J. (1992). Active shape models – smart snakes. In *Proc. of the 3rd British Machine Vision Conference 1992*, pages 266 – 275. Springer Verlag.

Edwards, G. J., Cootes, T. F., and Taylor, C. J. (1998). Face recognition using active appearance models. In Burkhardt, H. and Neumann, B., editors, *5th European Conference on Computer Vision*, volume LNCS-Series 1406–1607, pages 581–595, Freiburg, Germany. Springer-Verlag.

Felzenszwalb, P. and Huttenlocher, D. (2000). Efficient matching of pictorial structures. In *International Conference on Computer Vision and Pattern Recognition*, pages 66–73.

database ABC	dim [#]	aggressive	cheerful	intoxicated	nervous	neutral	tired	RR [%]	CL [%]	F_1 [%]
audio	276	87.3	65.7	52.6	63.1	61.0	76.3	69.4	66.9	68.1
audio opt.	92	90.4	71.4	40.0	66.7	70.4	81.1	73.7	68.8	71.2
video	1048	55.4	62.1	37.7	51.4	41.5	44.2	51.8	48.2	49.9
video opt.	156	59.4	65.3	47.5	65.2	60.2	58.8	61.1	58.4	59.7
av	1324	85.9	72.1	47.3	67.2	62.9	74.7	71.2	67.5	69.3
av ind. opt.	248	90.6	80.0	61.0	71.8	65.1	81.0	77.3	74.6	75.9
av comb. opt.	200	91.6	81.0	71.2	80.6	73.8	85.3	81.8	79.8	80.8

Table 4: Behavior confusions and f_1-measures by use of SVM in a 10-fold SCV, optimized audiovisual feature set.

Hanek, R. (2004). *Fitting Parametric Curve Models to Images Using Local Self-adapting Seperation Criteria*. PhD thesis, Department of Informatics, Technische Universität München.

Littlewort, G., Fasel, I., Bartlett, M. S., and Movellan, J. R. (2002). Fully automatic coding of basic expressions from video. Technical report.

Michel, P. and Kaliouby, R. E. (2003). Real time facial expression recognition in video using support vector machines. In *Fifth International Conference on Multimodal Interfaces*, pages 258–264, Vancouver.

Pantic, M. and Rothkrantz, L. J. M. (2003). Toward an affect-sensitive multimodal human-computer interaction. *Proceedings of the IEEE, Special Issue on human-computer multimodal interface*, 91(9):1370–1390.

Schuller, B., Mueller, R., Lang, M., and Rigoll, G. (2005). Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensembles. In *Proc. Interspeech 2005*, Lisboa, Portugal. ISCA.

Schuller, B. and Rigoll, G. (2006). Timing levels in segment-based speech emotion recognition. In *Proc. INTERSPEECH 2006*, Pittsburgh, USA. ISCA.

Schweiger, R., Bayerl, P., and Neumann, H. (2004). Neural architecture for temporal emotion classification. In *Affective Dialogue Systems 2004, LNAI 3068*, pages 49–52, Kloster Irsee. Elisabeth Andre et al (Hrsg.).

Sigal, L., Sclaroff, S., and Athitsos, V. (2000). Esitmation and prediction of evolving color distributions for skin segmentation under varying illumination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2000)*.

Viola, P. and Jones, M. J. (2004). Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154.

Wimmer, M., Pietzsch, S., Stulp, F., and Radig, B. (2007a). Learning robust objective functions with application to face model fitting. In *Proceedings of the 29th DAGM Symposium*, volume 1, pages 486–496, Heidelberg, Germany.

Wimmer, M., Radig, B., and Beetz, M. (2006). A person and context specific approach for skin color classification. In *Procedings of the 18th International Conference of Pattern Recognition (ICPR 2006)*, volume 2, pages 39–42, Los Alamitos, CA, USA. IEEE Computer Society.

Wimmer, M., Stulp, F., Pietzsch, S., and Radig, B. (2007b). Learning local objective functions for robust face model fitting. In *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*. to appear.