16th European Signal Processing Conference (EUSIPCO 2008), Lausanne, Switzerland, August 25-29, 2008, copyright by EURASIP

AUDIO-VISUAL EMOTION RECOGNITION USING AN EMOTION SPACE CONCEPT

Ittipan Kanluan, Michael Grimm, Kristian Kroschel

Universität Karlsruhe (TH), Institut für Nachrichtentechnik
Kaiserstr. 12, 76128 Karlsruhe, Germany
Email: michael.grimm@ieee.org

ABSTRACT

In this paper, we present novel methods for estimating spontaneously expressed emotions using audio-visual information. Emotions are described with three continuous-valued emotion primitives, namely valence, activation, and dominance in a 3D emotion space. We used prosodic and spectral features to represent the audio characteristics of the emotional speech. For the extraction of visual features, the 2-dimensional Discrete Cosine Transform (2D-DCT) was applied to blocks of a predefined size in facial images. Support Vector Machines (SVM) are used in their application for regression (Support Vector Regression, SVR) to estimate these 3 emotion primitives. The result showed that the emotion primitives activation and dominance can be best estimated with acoustic features, whereas the estimation of valence yields the best result when visual features are used.

Both monomodal emotion estimations were subsequently fused at a decision level by a weighted linear combination. The average estimation error of the fused result was 17.6% and 12.7% below the individual error of the acoustic and visual emotion recognition, respectively. The correlation between the emotion estimates and the manual reference was increased by 12.3% and 9.0%, respectively.

1. INTRODUCTION

Emotion recognition plays an important role in the research area of human-machine interaction. It allows a more natural and more human-like communication between humans and computer. This is particularly useful, for example, in the design of humanoid robots [1], where emotional intelligence is of great significance. The human sensory system uses multimodal analysis of multiple communication channels to interpret face-to-face interaction and to recognize another party's emotion. The psychological study [2] indicated that humans mainly rely on vocal intonation and facial expression to judge someone's emotional state. Hence, automatic emotion recognition systems should at least make use of these two modalities to achieve a reliable and robust result.

Based on a cross-cultural study, Ekman and Friesen postulated six basic emotions that can be displayed through unique facial expressions [3]: *happiness*, *sadness*, *anger*, *fear*, *surprise* and *disgust*. Most researches on emotion recognition so far have tried to classify human emotions in these six basic categories. Moreover, most of them concentrated on recognizing emotion either from speech [4] or from facial expression [5]. Relatively few of the existing works have been done in researching multimodal emotion recognition. Examples are the works of De Silva and Ng [6], and Chen et al. [7]. These works studied the effects of a combined detection of vocal and facial expressions of emotional states. Chen et al. [7] found that humans recognize some emotions better by audio information, and others better by video. They also showed that using both modalities makes it possible to achieve higher recognition rates than either modality alone. Recently, there has been an increasing interest in studying multimodal emotion recognition [8, 9]. Hoch et al. [8] used statistical characteristics of prosodic parameters and a set of Gabor wavelets to represent acoustic and visual features, respectively. The results from monomodal recognition were fused at a decision level by a weighted linear combination. This yields a better average recognition rate of 3.9% compared to the best monomodal classifier.

Most studies mentioned above used acted emotions performed by speakers. Moreover, all of them concentrated on classifying emotions into one of the basic emotion categories. This approach has many limitations because pure expressions of basic emotions are seldom elicited. Since humans generally show blends of emotional displays, the classification of human emotion into a single basic emotion category is not realistic [10].

In this paper, we propose a new way for authentic emotion recognition. Emotions are not classified into one of the emotion categories; they are estimated on a continuous-valued scale of three emotion attributes. We describe emotions using an emotion space concept as proposed by Kehrein [11]. This emotion space representation is in terms of three emotion primitives, namely *valence* (negative vs. positive nature of an emotion), *activation* (excitation level from calm to excited), and *dominance* (appearance of the person from weak to strong). Without loss of generality, they can be normalized to take values in the range of [-1,+1] each. Estimating emotions on a continuous-valued scale provides an essential framework for recognizing dynamic in emotions, tracking intensities in the course of time, and adapting to individual moods or personalities. As acoustic features, we use prosodic parameters and Mel Frequency Cepstral Coefficients (MFCCs). For feature extraction from the video signal, we apply the 2D-DCT to blocks of a predefined size in every image frame. DCT coefficients that are highly correlated with emotion primitives are retained as relevant visual features. Feature selection techniques are used to reduce the size of the feature vectors. Our previous work showed that Support Vector Regression (SVR) performs better than a Fuzzy *k*-Nearest Neighbor estimator or a rule-based Fuzzy Logic estimator in estimating emotion primitives in speech [12]. Hence, SVR is used as an estimator in this study. Another reason for choosing SVR is that it is based on a solid

This work has been supported by the German Science Foundation DFG within the Sonderforschungsbereich 588 "Humanoid Robots".

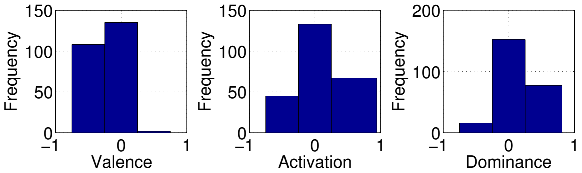

Figure 1: Histogram of emotion evaluation in VAM-Fusion-Audio database

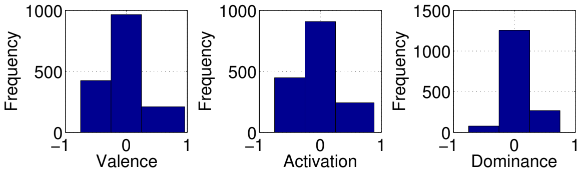

Figure 2: Histogram of emotion evaluation in VAM-Fusion-Faces database

theoretical framework to minimize the structural risk and not only the training error (empirical risk) [13]. The results of individual monomodal estimation are then fused at the decision level by a weighted linear combination. Our goal is to examine whether emotion recognition using both modalities yields better and more robust results than the one using only one modality. Furthermore, we analyze the relevance of each modality on the estimation of the individual emotion primitives.

The rest of this paper is organized as follows. Section 2 briefly introduces the data we used, and it also describes the emotion evaluation by human listeners. Section 3 describes monomodal emotion estimation. It firstly explains the feature extraction and feature selection and then presents the results of acoustic and visual emotion estimation. Section 4 presents the fusion step for both modalities and its results. Section 5 contains the conclusion and directions for the future work.

2. DATA AND EVALUATION

2.1 Database

The *VAM corpus* [14] was used. This audio-visual database was recorded from the German TV talk show “Vera am Mittag” in which guests mostly talk about their personal issues such as friendship problems and fatherhood questions in a spontaneous, affective and unscripted manner. For emotion recognition in the speech signal, the dialogues were segmented into utterances. The signals were sampled at 16 kHz and 16 bit resolution. Facial image sequences were taken from the video signal at the rate of 25 frames per second. For this study, we extracted a subset of corresponding audio and image files from the 12-hours-main corpus, which we call *VAM-Fusion*. This database thus consists of two modules, *VAM-Fusion-Audio* and *VAM-Fusion-Faces*. We used 245 utterances of 20 speakers in *VAM-Fusion-Audio* as the basis for the acoustic emotion estimation. For the visual emotion estimation, 1600 images from *VAM-Fusion-Faces* were used. For the audio-visual emotion estimation purpose, we chose only sentences in *VAM-Fusion-Audio* whose corresponding images in *VAM-Fusion-Faces* were evaluated. Finally, we used 234 sentences and 1600 images from *VAM-Fusion-Audio* and *VAM-Fusion-Faces*, respectively.

Fig. 1 and 2 show the histogram of the emotions contained in the databases *VAM-Fusion-Audio* and *VAM-Fusion-Faces* respectively. The emotion attested by human listeners (see section 2.2) was taken as the reference for the automatic recognition, since assessments by the speakers themselves were not available.

2.2 Emotion evaluation

For evaluation we used an icon-based method based on *Self Assessment Manikins (SAM)* [15] that yields one reference value $x_n^{(i)}$ for each primitive $i \in \{valence, activation, dominance\}$. For each acoustic utterance and each facial image n, the evaluators were asked to select one of five given iconic images that best describes each emotion primitive. These five iconic images were oriented from negative to positive (*valence*), from calm to excited (*activation*), and from weak to strong (*dominance*). This evaluation was subsequently mapped to a scale of [-1,+1] for each primitive. The individual evaluator ratings were averaged using confidence scores as described in [15].

Table 1: Standard deviation $\bar{\sigma}$ and correlation coefficient $\bar{r}$ for the emotion primitives evaluation of the VAM-Fusion database, averaged over all speakers and all sentences

VAM-Fusion-Audio	Valence	Activation	Dominance
Std. deviation $\bar{\sigma}$	0.29	0.35	0.31
Correlation coeff. $\bar{r}$	0.60	0.81	0.72
VAM-Fusion-Faces	**Valence**	**Activation**	**Dominance**
Std. deviation $\bar{\sigma}$	0.37	0.44	0.48
Correlation coeff. $\bar{r}$	0.59	0.63	0.40

The average standard deviation in the evaluation of both utterance and facial image was calculated as shown in Tab. 1. The evaluators showed moderate to high inter-evaluator agreement as can be calculated from Pearson’s correlation coefficient, c.f. Tab. 1. The mean correlation between the evaluators for the acoustic evaluation was 0.60, 0.81, and 0.72 for *valence*, *activation*, and *dominance*, respectively. Thus, *valence* was significantly more difficult to evaluate than *activation* or *dominance*. For visual evaluation, the mean correlation between the evaluators was 0.59, 0.63, and 0.40, respectively. Hence, *dominance* was notably harder to evaluate from facial expression than *valence* or *activation*. In general, the speech signals were evaluated on a significantly higher inter-evaluator agreement than the facial expressions. One can notice from Fig. 1 and Fig. 2 that the emotion evaluation for all the three emotion primitives is mainly negative. This is due to the topic of the discussion in the talk show from which the *VAM corpus* was recorded.

3. MONOMODAL EMOTION ESTIMATION

3.1 Acoustic emotion estimation

3.1.1 Feature extraction

In accordance with other studies on automatic emotion recognition we extracted prosodic features from the fundamental frequency (pitch) and the energy contours of the speech signals. The first and the second derivatives of these contours were also used. From these signals we calculated the following statistical characteristics: mean value, standard deviation, median, maximum, minimum, 25% and 75% quantiles, difference between the maximum and minimum, and difference between the quartiles. For the temporal characteristics, we used speaking rate, pause to speech ratio, mean and standard deviation of the speech duration, and mean and standard deviation of the pause duration. In addition we also used spectral characteristics in 13 subbands derived from the MFCCs. Totally, 137 acoustic features were extracted. They were normalized to the range [0,1].

3.1.2 *Feature selection*

To reduce the large number of acoustic features, we used the Sequential Forward Floating Search (SFFS) technique for feature selection. We found that, for each of the emotion primitives, using 20 features was sufficient, and adding more features hardly improved the results. In addition, SFFS gave slightly better results compared to Principal Component Analysis (PCA).

3.1.3 *Emotion estimation*

We used Support Vector Regression (SVR) to estimate continuous values $x_n^{(i)}$ of emotion primitives i, $i \in \{$*valence*, *activation*, *dominance*$\}$. SVR is a regression method based on Support Vector Machines (SVM). It tries to find the optimal regression hyperplane so that most training samples lie within an ε-margin around this hyperplane [13]. In this study, we performed a non-linear regression by applying the *Kernel trick*, i.e., to replace the inner product in the solution by a non-linear kernel function. We chose a radial basis function (RBF) with $\sigma = 3.5$ as a kernel function. In our previous work, the RBF was proven to be the best kernel function over polynomial kernel and linear kernel in estimating emotion primitives in speech [12]. The design parameters of the SVR, C and ε, were selected using a grid search on a logarithmic scale and a second, fine-grained search in the best region. We chose $C = 10$ and $\varepsilon = 0.2$.

3.1.4 *Result of acoustic estimation*

We performed a 10-fold cross-validation and calculated the mean linear error and correlation between the estimates and the references ($N = 234$) for each emotion primitive separately. The correlation shows the accuracy in the tendency of the estimates. The mean linear error was 0.13, 0.16, and 0.14 for *valence*, *activation*, and *dominance*, respectively. This showed that all 3 emotion primitives can be similarly well estimated within a small range of error. However, the correlation between the estimates and the reference was significantly different for the individual emotion primitives. It was 0.53, 0.82, and 0.78 for *valence*, *activation*, and *dominance*, respectively. After performing the statistical significance test (p-value), the correlations for activation and dominance showed statistical significance with small p-values ($p < 10^{-5}$), whereas the correlation for valence was not statistically significant ($p > 0.05$). Thus, the results implied very good recognition results for *activation* and *dominance*, and only moderate recognition result for *valence*.

3.2 Visual emotion estimation

3.2.1 *Pre-processing*

Firstly, the face was detected in an image grabbed from the video stream. We used the real-time face detection algorithm by Viola and Jones [16]. Each facial image was converted to grayscale and segmented into two subimages, the upper and the lower face, respectively. The eyes region was determined by locating the eye positions within the upper face image and scaling the relevant image section to a size of 100×150 pixels. The lip region was determined by locating the mouth within the lower face image and scaling the relevant image section to a size of 75×75 pixels. Normalization was applied, since the size of the face was not the same in all images. Fig. 3 shows the extraction of these regions of interest from the face image.

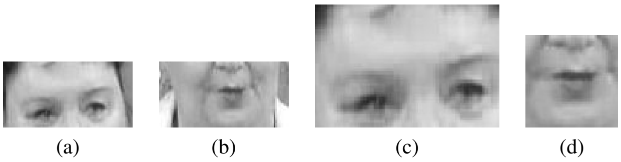

Figure 3: Segmentation of the face image into eyes region and lip region: (a) upper face, (b) lower face, (c) normalized eyes region, (d) normalized lip region

Note that almost all of the facial images in our database showed a frontal, upright pose. Illumination normalization was not needed because all images were under the same lighting condition.

3.2.2 *Feature extraction*

We used the 2-dimensional Discrete Cosine Transform (2D-DCT) for the feature extraction step. There are four established types of DCT, i.e., DCT-I, DCT-II, DCT-III, and DCT-IV. The DCT-II is more widely applied in signal coding because it is asymptotically equivalent to the Karhunen-Loève Transform (KLT) for Markov-1 signals with a correlation coefficient that is close to one [17]. The DCT-II is often simply referred to as “the DCT ”. The 2D $M \times N$ DCT is defined as follows [17]:

$$C(u,v) = \alpha(u)\alpha(v) \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x,y) \times \cos\left(\frac{\pi(2x+1)u}{2M}\right) \cos\left(\frac{\pi(2y+1)v}{2N}\right) \quad (1)$$

where

$$\alpha(u) = \begin{cases} \sqrt{\frac{1}{M}} & \text{for } u = 0 \\ \sqrt{\frac{2}{M}} & \text{for } u = 1,2,...,M-1 \end{cases}$$

$$\alpha(v) = \begin{cases} \sqrt{\frac{1}{N}} & \text{for } v = 0 \\ \sqrt{\frac{2}{N}} & \text{for } v = 1,2,...,N-1 \end{cases}$$

and $f(x,y)$ is the gray value at position (x,y) in the normalized facial image region. We applied the 2D-DCT to $M \times N$ blocks of pixels in the eyes region and lip region images in our database. Our goal was to find the DCT coefficients that were relevant to the variation of each emotion primitive in the image sequence. In this study, we chose $M = N$ and considered four cases of block size, i.e., 8×8, 16×16, 32×32, and 50×50 pixels. For each case, block overlapping of 50% in pixels was also studied. After applying the

Table 2: Emotion primitives estimation results for the eyes and the lip region: mean error and correlation coefficient

Mean error	Valence	Activation	Dominance
Eyes region	0.18	0.19	0.13
Lip region	0.18	0.19	0.14
Correlation coeff.	Valence	Activation	Dominance
Eyes region	0.57	0.58	0.57
Lip region	0.58	0.62	0.53

2D-DCT to $N \times N$ blocks in the images, the number of features was $N \times N \times$ *number of blocks in image*. For example, for feature extraction of the eyes region images (100×150 pixels) using non-overlapping 8×8 blocks of pixels, we have 8×8×247 = 15,808 features.

3.2.3 *Feature selection*

We used SVR-SFFS technique for feature selection. Firstly, the DCT coefficients between 10% and 90% quantile were normalized to a scale of [0,1]. Then, they were sorted in descending order according to their correlation to the manual labels of the emotion primitives. We performed SVR-SFFS by taking the DCT coefficients that had the largest absolute value of correlation coefficient first, and so on. Fig. 4 shows the mean linear error and correlation coefficient between the estimates and the manual reference for each emotion primitive as a function of the feature set. For each emotion primitive, the number of features that were retained for the estimation was the number that yields the lowest mean linear error. Hence, we used 167, 113, and 134 coefficients for *valence*, *activation*, and *dominance*, respectively. Adding more features hardly improves the estimation results since one may actually be representing more irrelevant information. For the feature extraction of the eyes region image using non-overlapping 50×50 blocks of pixels, we extracted 136, 126, and 137 coefficients, respectively. We found that these relevant features are DCT coefficients from both low-frequency and high-frequency bands of various blocks within the facial image regions.

3.2.4 *Emotion estimation*

We also used SVR with a RBF kernel for visual emotion estimation. Using the RBF kernel has shown the best performance over other kernels in many applications such as emotion recognition in speech [12]. We chose $\sigma = 2.5$, since this was proven to give the best performance for visual emotion estimation. For the SVR parameters, we selected $C = 10$ and $\varepsilon = 0.2$.

3.2.5 *Result of visual estimation*

The results were achieved using 10-fold cross-validation (N = 1600). We found that features which were extracted by applying the 2D-DCT to non-overlapping 50×50 blocks of pixels yield the best estimation result for both the eyes region and lip region. This is due to the size of block which is large enough to cover the characteristics of a facial expression in the subregion of images. Tab. 2 shows the mean linear error and correlation for the eyes and lip region, for each emotion primitive separately. It can be seen that both regions of interest can be similarly well estimated with a small error in the range of 0.13 to 0.19. The correlation for each of the three emotion primitives does not differ significantly. It was in the range of 0.57 to 0.58 and 0.53 to 0.62 for the eyes region and the lip region, respectively. This tendency of the estimates for both regions was also statistically significant with p-values $< 10^{-5}$. Thus, the results implied good recognition results for all three emotion primitives.

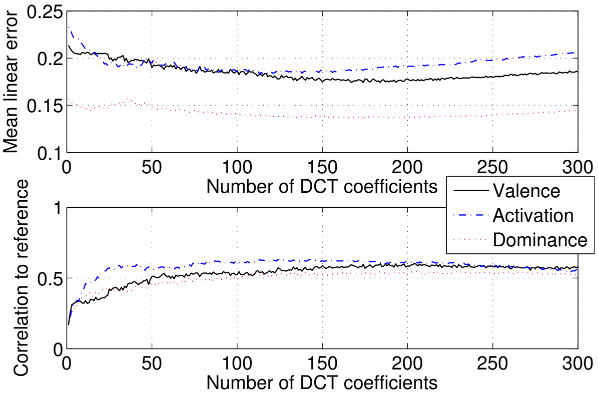

Figure 4: Mean error and correlation coefficient in the application of SVR-SFFS to the features extracted by non-overlapping 50×50 blocks of pixels from the lip region images

4. FUSION

We performed the fusion of the acoustic information and the facial expression at the decision level. The synchronisation aspect between the acoustic and the visual information was considered, i.e., we fused those two modalities by combining the results of the acoustic estimation of the selected utterances with the results of the visual estimation of the facial images that correspond to these utterances. Since the facial expression/emotion can be changed during an utterance, the *Maximum Likelihood Estimator (MLE)* was used to combine the visual emotion estimation results corresponding to a given utterance. Thus, the number of visual emotion estimation results for all utterances was reduced from 1600 to 234 values.

4.1 Fusion architecture

The reduced number of visual estimation results was fused with the results of acoustic estimation of 234 sentences using a weighted linear combination as follows:

$$\hat{x}_n^{Av,(i)} = w^{(i)} \hat{x}_n^{Vis,(i)} + (1 - w^{(i)}) \hat{x}_n^{Ac,(i)} \tag{2}$$

where $\hat{x}^{Av}$, $\hat{x}^{Vis}$ and $\hat{x}^{Ac}$ is the fusion result, the visual estimation result, and the acoustic estimation result for sentence n, respectively. We define w as the visual weighting factor with $i \in \{valence, activation, dominance\}$.

4.2 Result of the fusion

We explored the influence of different weighting factors w of the modalities for each emotion primitive in the fusion process separately, ranging from pure acoustic estimation (w

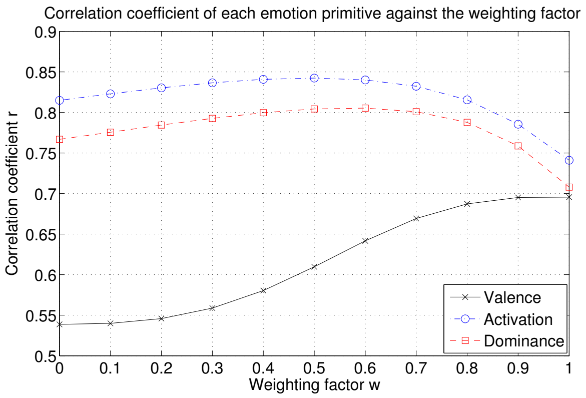

Figure 5: Correlation of each emotion primitive against the visual weighting factor *w*

= 0.0) to pure facial expression analysis (w = 1.0). Fig. 5 shows the results of fusion comparing the correlation of each emotion primitive to its reference against the used weighting factor w. The fusion for *valence* yields the best estimation result when only visual estimates were used (w = 1.0). In this case, the correlation coefficient was 0.70. For *activation* and *dominance*, the fusion results were best when both visual and acoustic estimates are combined with a weighting factor w = 0.5 and w = 0.6, respectively. The correlation coefficient was 0.84 and 0.80 for *activation* and *dominance*, respectively. The mean linear error was 0.14, 0.12, and 0.09 for *valence*, *activation*, and *dominance*, respectively. We found that both, the eyes region and the lip region yield equally good recognition results. The fusion results implied very good recognition results for all three emotion primitives with a low error in the range of 0.09 to 0.14 and a high correlation in the range of 0.70 to 0.84.

5. CONCLUSION AND OUTLOOK

We investigated the continuous-valued estimation of three emotion primitives, namely *valence*, *activation*, and *dominance*, using acoustic and visual features. The acoustic features were extracted from the prosody and the spectrum of spontaneous speech signals. For visual feature extraction, we applied the 2D-DCT to blocks of a predefined size in facial images. The emotion primitives are estimated with a small error of 0.13 to 0.19, where the range of values was [-1,+1], for both modalities. The estimation of *activation* and *dominance* using acoustic features shows a high correlation between the estimates and the manual labels, 0.82 and 0.78 respectively, to the reference, whereas *valence* can be better estimated using visual features with a correlation of 0.57. The fusion of these two modalities was executed at the decision level using a weighted linear combination. For the mean error, the results showed an average performance gain of 17.6% and 12.7% over the individual acoustic and visual emotion estimation, respectively. For correlation, the average performance gain was 12.3% and 9.0%, respectively. The fusion results imply that *valence* can be best estimated by the visual information alone. For the estimation of *activation* and *dominance*, combining both modalities with the visual weighting factor w = 0.5 and w = 0.6, respectively, achieves the best recognition results.

In future work, designing a real-time system using the algorithms reported here should be investigated. Furthermore, the adaptive adjustment of the weighting factor w in the fusion process should be studied. This will allow us to build a more robust audio-visual emotion recognition system where the weighting factor could be dynamically changed according to several parameters, e.g. the SNR of the data.

REFERENCES

[1] "Humanoid Robots – Learning and Cooperating Multimodal Robots," Collaborative Research Center of the Deutsche Forschungsgemeinschaft, http://www.sfb588.uni-karlsruhe.de/, 2001.

[2] A. Mehrabian, "Communication without words," *Psychol. Today*, vol. 2, no. 4, pp. 53–56, 1968.

[3] P. Ekman and W. Friesen, "Constants across cultures in the Face and Emotion," *Journal of Personality and Social Psychology*, vol. 17, no. 2, pp. 124–129, 1971.

[4] O. Kwon, K. Chan, et al., "Emotion recognition by speech signals," in *Eurospeech*, Geneva, Sep. 2003.

[5] M. J. Lyons, J. Budynek, et al., "Classifying facial attributes using a 2-D Gabor wavelet representation and discriminant analysis," in *Proc. FG*, March 2000.

[6] L. C. De Silva and P. C. Ng, "Bimodal emotion recognition," in *Proc. FG*, pp. 332–335, March 2000.

[7] L. Chen, T. S. Huang, et al., "Multimodal human emotion/expression recognition," in *Proc. FG*, pp. 396–401, Apr. 1998.

[8] S. Hoch, F. Althoff, et al., "Bimodal fusion of emotional data in automotive environment," in *Proc. ICASSP*, vol. 2, pp. 1085–1088, 2005.

[9] Y. Wang and L. Guan, "Recognizing human emotion from audiovisual information," in *Proc. ICASSP* , vol. 2, pp. 1125–1128, 2005.

[10] M. Pantic and L. J. M. Rothkrantz, "Toward an affect-sensitive multimodal human-computer interaction," in *Proc. IEEE*, vol. 91, pp. 1370–1390, Sep. 2003.

[11] R. Kehrein, "The prosody of authentic emotions," in *Proc. Speech Prosody Conf.*, pp. 423–426, 2002.

[12] M. Grimm, K. Kroschel, and S. Narayanan, "Support Vector Regression for automatic recognition of spontaneous emotion in speech," in *Proc. ICASSP*, vol. 4, pp. IV-1085–IV-1088, 2007.

[13] V. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer, 1995.

[14] M. Grimm, K. Kroschel, and S. Narayanan "The Vera am Mittag German audio-visual emotional speech database," Submitted to the IEEE Int. Conf. on Multimedia & Expo (ICME), Hannover, Germany, 2008

[15] M. Grimm and K. Kroschel, "Evaluation of natural emotions using self assessment manikins," in *Proc. of IEEE ASRU Workshop*, pp. 381–385, 2005.

[16] P. Viola and M. Jones, "Robust real-time object detection," in *Proc. of 2nd IEEE Workshop on Statistical and Computational Theories of Vision*, pp. 1–25, 2001.

[17] K. R. Rao and P. Yip, *Discrete Cosine Transform: Algorithms, Advantages, Applications*. Boston, MA: Academic Press, 1990.