RECOGNIZING EMOTIONS IN SPONTANEOUS FACIAL EXPRESSIONS

Michael Grimm, Dhrubabrata Ghosh Dastidar, and Kristian Kroschel

Institut für Nachrichtentechnik
Universität Karlsruhe (TH), Germany
grimm@int.uni-karlsruhe.de

ABSTRACT

In this paper we present a method for classifying emotions in spontaneous facial expressions of both active and inactive speakers in spoken dialogues. Evaluation and classification was performed for emotion categories (happiness, sadness, anger, fear, surprise, disgust, neutral) and emotion space classes (3 classes for valence and activation, respectively). In addition, continuous values of the emotion space attributes were estimated.
For feature extraction, a novel combination of multi-scale, multi-orientation Gabor filtering and Principal Component Analysis was used. For classification, an Artificial Neural Net was used, with a fuzzy logic extension for the estimation of the continuous-valued emotion space attributes.
The maximum average recognition rate for emotion category and for emotion space classification was 72.9% and 80.1%, respectively. The mean error for continuous-valued emotion primitives estimation was 0.3, when the range of values was [-1,+1]. A FACS-adapted extension was also introduced, defining a two-stage description of expressive meta-features in the face, such as *open eyes* vs. *closed eyes* as a stage 1 meta-feature, and *open mouth, smiling* vs. *open mouth, not smiling* as a stage 2 meta feature. Using this 2-stage classification method, an average recognition rate between 81.7% and 99.1% was achieved for the individual classifications. It was found that, although we were using spontaneous instead of posed facial expressions, our results almost achieved the recognition rates reported in the literature.

1. INTRODUCTION

In recent years there has been a growing interest in improving the interaction between human operators and intelligent machines such as computer tutors or service robots. To achieve a natural, human-like interaction, the dialogue strategy should be adapted to the emotional state of the user. Emotions are a vital component of any person and can be expressed by different modalities, the most common being speech, head pose, gestures and facial expressions. In this paper we will concentrate on the automatic classification of emotions from spontaneous facial expressions.
Facial expression analysis dates back to the nineteenth century, when in 1872 Darwin demonstrated the universality of facial expressions in man and animals [1]. The automatic, robust recognition of emotions from facial expressions is still an unsolved problem, in particular when spontaneous facial expressions are used. Fasel and Luettin [2] give an excellent survey on the research in the area of facial expression recognition.
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 of the works on automatic recognition of emotions from facial expressions so far have concentrated on classifying the emotions into these basic emotions [4, 5, 6]. Padgett and Cottrell [4] used Principal Component Analysis and Artificial Neural Networks (ANN) to classify these emotions in facial images of Ekman's "Pictures of Facial Affect" collection. A recognition rate of 84% was reported. Lyons *et al.* [5] achieved a recognition rate of 92% on a person-dependent classification task using Linear Discriminant Analysis and a dataset of 9 Japanese persons who had posed 3 or 4 examples of each of the basic emotions (JAFAE database [7]). Bartlett *et al.* [6] used facial pictures of the Cohn-Kanade database [8]. They report a recognition rate of 88% using Gabor filtering and Support Vector Machines.
A different approach to facial expression recognition is using the Facial Action Coding System (FACS) [9]. It defines all possible facial movements in terms of component actions motivated by the biology of the facial muscles. Donato *et al.* [10] compared several techniques, including Independent Component Analysis and Gabor wavelet representations for recognizing facial expressions by automated FACS encoding. They achieved a recognition rate of 95%. Tian *et al.* [11] automatically recognized 7 Action Units (AU) in the upper face, 11 AUs in the lower face and several combinations of them. Tested on the Cohn-Kanade database, they also report a recognition rate of 95%. The major problem in these approaches is the need of highly trained experts to manually label the facial expressions in the FACS as a reference.
All of the above mentioned methods use posed facial expressions which show, in general, exaggerated emotions. A promising method of obtaining spontaneous emotions is described by Sebe *et al.* [12]. They installed a hidden camera in a video kiosk which played recent movie trailers. Another naturalistic emotion database is the Belfast database [13] that was used in the recent study of Ioannou *et al.* [14].
In our study we use authentic emotions expressed by guests in a talk show broadcasted on TV. Motivated by promising results in the automatic recognition of emotions from acoustic features in the speech signal [15], we use the emotion space concept for describing the emotions in facial expressions. One powerful emotion space representation is in terms of the three emotional attributes ("primitives") namely *valence* (positive vs. negative), *activation* (excitation level high vs. low), and *dominance* (apparent strength or weakness of the speaker) [16]. In this study we classify the facial expressions into the six Ekman emotion categories and into three subspaces for each primitive. In addition we also estimate

This work was supported by grants of the German Academic Exchange Service (DAAD) and the Collaborative Research Center (SFB) 588 "Humanoid Robots" of the Deutsche Forschungsgemeinschaft (DFG).

continuous values of these emotion primitives.

As feature extraction technique we use a deformation-based approach based on 18 Gabor wavelet filters for 3 different scales and 6 different orientations. Principal Component Analysis (PCA) is performed on them for dimensionality reduction, and then these features are classified using Artificial Neural Networks (ANN).

The novel aspects in this paper include the following: A new database containing authentic facial expressions is presented and used for classification. An image-based evaluation method using Self Assessment Manikins (SAMs) is used for generating the emotion space reference of the facial expressions. Features extracted from Gabor filtering of the facial images are used for emotion space classification, and a novel combination of PCA and Gabor filtering is applied. A neuro-fuzzy method is used for contiuous-valued emotion estimation. Finally, a FACS-adapted approach is introduced, defining easily observable meta-features, such as *open eyes* vs. *closed eyes*, and *open mouth smiling* vs. *open mouth non-smiling*.

The rest of the paper is organized as follows. Section 2 describes the database that we used to test the proposed algorithms. Section 3 describes the methods used for pre-processing and feature extraction as well as the classifier used. Section 4 presents details of the classification into the emotion categories and into the subclasses defined in the emotion space. It also contains the details of the neuro-fuzzy approach for continuous-valued emotion attributes estimation. Section 5 proposes the use of meta-features for classification. Section 6 draws some conclusions and outlines future work.

2. DATA

2.1. Database

For this study we used the VAM Corpus. It was recorded from a German talk show on free TV called "Vera am Mittag". The speakers mostly discuss personal problems or family issues in a spontaneous unscripted fashion, showing a wide range of spontaneous emotions. The speech part of the VAM Corpus was initially used for emotion detection in speech signals [17].

Facial image sequences were extracted from the video signal at the rate of 25 frames per second and a resolution of 352X288 pixels. Note that the extracted images contain both faces of speaking and non-speaking persons in contrast to most other databases mentioned in Section 1. Of the initial database of 47 speakers we use only those which have few occlusions and at least two different emotion expressions. Thus our database contains 20 speakers (14 female, 6 male) with an average of 94 images per speaker and a total of 1872 images.

2.2. Evaluation

The images were randomly divided into different sets and evaluated for both the emotion categories as well as for emotion primitives *valence, activation*, and *dominance*. Since not all evaluators assessed all sets, the number of evaluators varied from 9 to 34.

2.2.1. Emotion category evaluation

For emotion category evaluation, the evaluators were asked to choose *neutral (N)* or one of the six basic emotions *happiness (H), sadness (Sa), anger (A), fear (F), surprise (Su)* and *disgust (D)* as described by Ekman [3]. This selection is called *primary emotion*

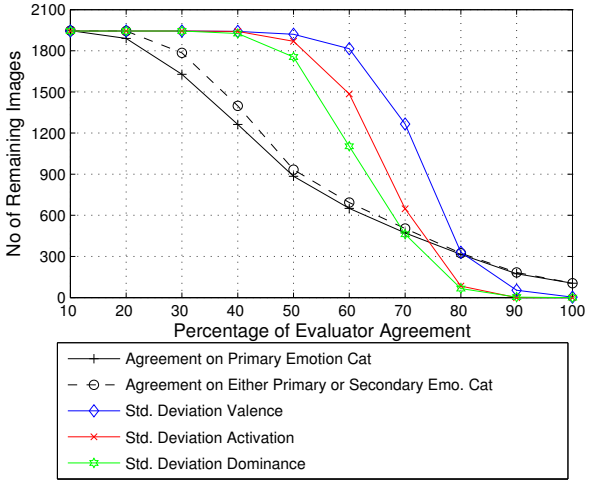

Fig. 1. Number of remaining sentences as a function of the required evaluator agreement for the VAM database, both for emotion category and emotion space evaluation.

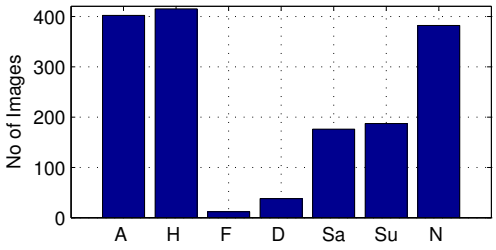

Fig. 2. Distribution of the emotions in the VAM database into the six basic emotion categories *anger (A), happiness (H), fear (F), disgust (D), sadness (Sa), surprise (Su)* and *neutral (N)*.

rating. They were also asked to choose a *secondary emotion* from the same set of emotions, if it was necessary to fully describe the emotional expression in the face by a mixture of two basic emotions. The evaluator agreement was calculated to decide which images to keep in the database and which ones to discard.

Fig. 1 shows the number of remaining images as a function of the evaluator agreement for this evaluation task. In general, the evaluator agreement on emotion categories was poor, since a majority vote meant discarding more than half of the database. We decided to require an agreement of at least 30%, which is still an agreement of at least 5 evaluators in most cases. The agreement on either the primary or the secondary emotion category was higher than on the primary only, however this is a trade-off between a higher number of images and a more reliable reference. Since we used a rather low agreement level, we decided for choosing primary emotion category agreement only. Thus the database reduced to a total of 1612 images of 20 speakers and an average of 81 images per speaker. The emotion category distribution is displayed in Fig. 2. Finally the number of images in each class was equalized resulting in a total of 764 images and a smaller number of images per

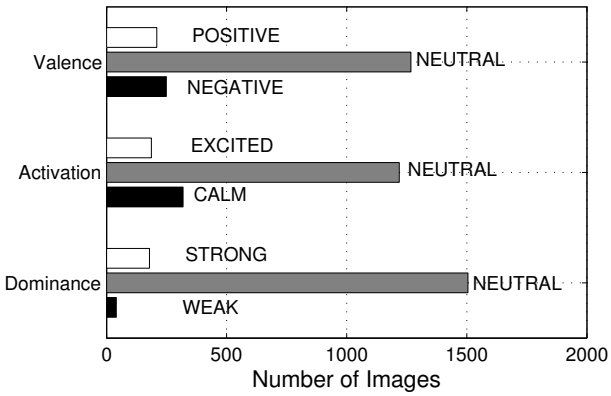

Fig. 3. Distribution of the emotion primitives in the VAM database into three different classes each.

speaker (min 27, max 72). Note that due to the naturalness of the dialogues not all emotions are expressed by all speakers. In the following, fear is not considered since it was present in too few images. The remaining database consists of 186 *anger*, 172 *happiness*, 90 *sad*, 110 *surprise*, 10 *disgust*, 0 *fear* and 196 *neutral* images.

2.2.2. Emotion space evaluation

The facial images were also evaluated in terms of the emotion space attributes *valence, activation* and *dominance*. A text-free method using five manikins (Self Assessment Manikins, SAMs) for each attribute was chosen [18]. The evaluators' selections were mapped to a scale of [-1, 1]. The average of the ratings was taken as a continuous-valued reference of the emotion position within the emotion space. The ratio between the standard deviation and the maximum possible standard deviation was chosen as an agreement measure.

Fig. 1 shows the number of remaining images as a function of the evaluator agreement also for this evaluation task. In general the agreement was significantly higher than on the emotion category evaluation task. Since it was decided to require the same level of agreement as above, we kept all images in the database. The resulting standard deviation and average correlation of the evaluators are given in Tab. 1. The values are worse than for evaluation of emotions in speech [18], but there is still a moderate positive correlation between the evaluators.

We discretized the continuous values into three classes for each of the primitives *valence* (→ *negative, neutral, positive*), *activation* (→ *calm, neutral, excited*) and *dominance* (→ *weak, neutral, strong*). The distribution of these classes is shown in Fig. 3. It was found that dominance is most difficult to assess in facial expressions since most of the evaluators have assessed neutral values for the dominance of the speaker. Thus for the facial expression recognition we restrict to classifying *valence* and *activation*.

To ensure images in at least two of the classes, 17 speakers for *valence* and 18 speakers for *activation* were chosen. A minimum of 15 images per speaker to a maximum of 78 images per speaker was available. The remaining database for this classification task consists of 141, 298, and 226 images for *positive, neutral*, and *negative* (*valence*), respectively, and it consists of 129, 311, and 242 images for *calm, neutral*, and *excited* (*activation*), respectively.

For the estimation of the continuous values of the emotion primitives, a neuro-fuzzy approach is used as described in Sec. 4.2.2. The reference values for this task were generated by assigning membership grades of the emotion primitives classes. These membership grades were obtained by fuzzifying the crisp values of the average evaluator ratings.

Table 1. *Standard deviation and correlation for valence, activation, and dominance in emotion evaluation.*

	Valence	Activation	Dominance
Standard Deviation	0.37	0.44	0.48
Correlation	0.45	0.53	0.57

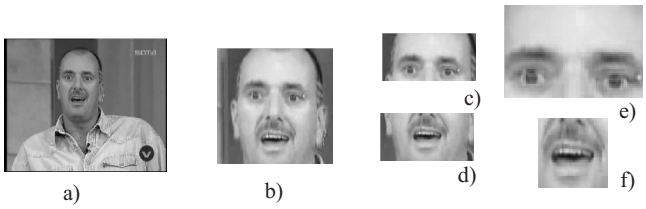

Fig. 4. Segmentation of the face into the eye and lip regions: a) original image (scaled 50%), b) detected face, c) upper face, d) lower face, e) normalized eye region, f) normalized lip region.

3. METHOD

The proposed method for recognizing emotions from spontaneous facial expressions consists of three steps: pre-processing, feature extraction, and classification. Each of them will be described in the following subsections.

3.1. Pre-processing

First of all, the face has to be detected in an image grabbed from the video stream. We use the real-time face detection algorithm by Viola and Jones [19]. The facial image is converted to grayscale and segmented into two subimages, the upper and the lower face, respectively. The eye region is determined by locating the eye positions within the upper face image [20] and scaling the relevant image section to a size of 150X100 pixels. The lip region is determined by locating the mouth within the lower face image and scaling the relevant image section to a size of 75X75 pixels. Normalization is applied since the size of the face is not the same in all images. Fig. 4 shows the extraction of these regions of interest for a sample image.

3.2. Feature Extraction

The feature extraction step consists of multi-scale, multi-orientation filtering of the region of interest images and a PCA for dimensionality reduction.

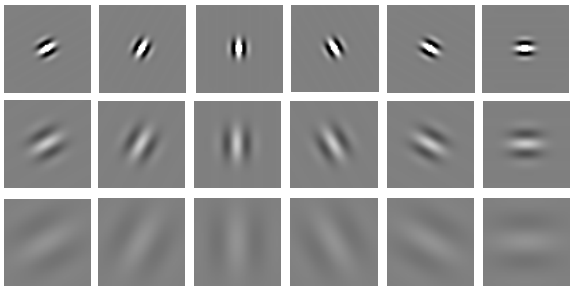

Fig. 5. Gabor filters used for feature extraction, where the row indicates the scale, and the column indicates the orientation of each filter.

3.2.1. *Gabor Filtering*

Gabor filtering has been shown to be a promising method for deformation-based facial expression analysis [5, 7, 21, 22, 23]. It has been found that Gabor wavelet-based features are relatively robust to illumination changes and head movement due to multiple resolution and multiple orientation filtering. A 2D Gabor filter $\Phi(\mathbf{k}, \mathbf{x})$ is defined as a Gaussian low-pass filter modulated by a plane wave,

$$\Phi(\mathbf{k}, \mathbf{x}) = \frac{|\mathbf{k}|^2}{\sigma^2} \exp\left(-\frac{|\mathbf{k}|^2 |\mathbf{x}|^2}{2\sigma^2}\right) \left(\exp(i\mathbf{k}^T\mathbf{x}) - \exp\left(-\frac{\sigma^2}{2}\right)\right), \tag{1}$$

where $\mathbf{x}$ represents the spatial localization and the wave vector $\mathbf{k} = (k\cos\theta, k\sin\theta)^T$ represents the translation and orientation of the tuned filter in the frequency domain [5]. The term $\exp\left(-\sigma^2/2\right)$ is subtracted to make the filters less sensitive to the absolute illumination level. In our approach the filters are modulated to three frequencies $k \in \{\pi/4, \pi/8, \pi/16\}$ and six orientations $\theta \in \{\pi/6, \pi/3, \pi/2, 2\pi/3, 5\pi/6, \pi\}$. The Gabor wavelet outputs are generated by convolving the region of interest images with the bank of 18 tuned Gabor filters shown in Fig. 5.

3.2.2. *Principal Component Analysis*

After Gabor filtering we have 18 filter outputs per region of interest. With the normalized image sizes of 150X100 and 75X75 pixels for the eye and the lip region, respectively, the filtering operation yields a total of $18 \cdot 150 \cdot 100 + 18 \cdot 75 \cdot 75 = 371{,}250$ features per image. To reduce this large number of features we use Principal Component Analysis (PCA) [24]. The PCA is applied to each region of interest for each orientation and each scale separately on a speaker-dependent basis.

The number of basis images (Eigenfaces) used is determined by the reconstruction error. Fig. 6 and Fig. 7 shows the reconstruction error for the eye and the lip region, respectively, as a function of the number of PCA coefficients used. It can be observed that for both the eye region and the lip region, a different number of coefficients is necessary for different scales. For the eye region we chose 30, 20, and 10 coefficients for each orientation of the scales $k = \pi/2, k = \pi/4$, and $k = \pi/8$, respectively, to achieve a reconstruction error below 1.2%. For the lip region we chose 20, 10, and 5 coefficients for each orientation of the scales $k = \pi/2, k = \pi/4$, and $k = \pi/8$, respectively, to achieve a reconstruction error below 1.5%. Thus 360 coefficients for the eye region and 210 coefficients for the lip region are calculated. These 570 features are then passed to the classifier.

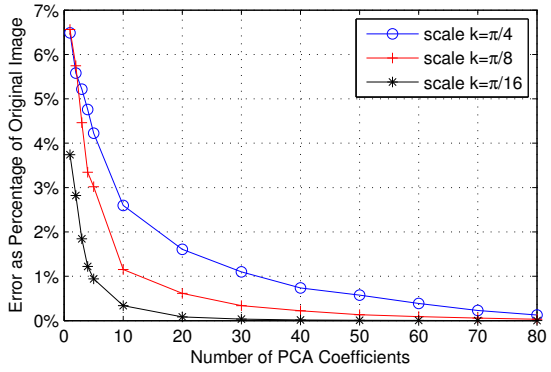

Fig. 6. Mean reconstruction error for all the orientations of each scale for the eye region.

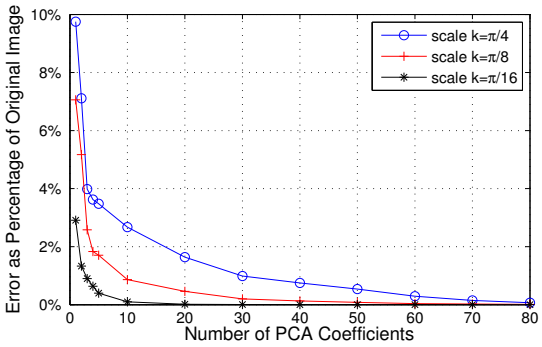

Fig. 7. Mean reconstruction error for all the orientations of each scale for the lip region.

3.3. Emotion Classification

An Artificial Neural Net (ANN) is used for the classification. A two-layer perceptron with one hidden layer in Feed-Forward topology is chosen [24]. Adaption of the weights is achieved using the Back-Propagation algorithm.

The classification is done for the lip region and the eye region separately, and with both taken together. It is done for the emotion categories as well as for the emotion space. At the output, seven neurons are used for the emotion categories. In the case of the emotion space classification, three output neurons for each of the attributes *valence* and *activation* are used. Due to the small number of images per speaker, testing was done based on leave one out (LOO) cross validation. The training set was used to train the net 10 times with an adaptation constant $\mu^{(i)} = 0.7 \cdot (5/6)^{i-1}$ in the ith iteration. Testing was done with the LOO test image.

4. RESULTS

In the following, classification results for both the emotion categories and the emotion subspace classes are provided. In each case the classification results are reported for the eye region and the lip region separately as well as for their combination.

4.1. Emotion category classification

The results for the emotion category classification are shown in Table 2 on page 8. The recognition of the six emotions was well above chance level. It can be observed that the maximum average recognition rate of 72.9% is achieved using the eye region alone. With an accuracy of 76.7%, *sadness* was best classified. The worst classification accuracy, on average 57.6%, was observed when using the lip region only. This is due to the poor accuracy for the classes *disgust* and, surprisingly, *neutral*, which was approx. 40%. The combination of the eye and lip regions results in an average accuracy of 68.2%, being in between the results of the two regions when taken separately. This might be due to the fact that the database was created from spoken dialogues, and thus the emotion in the lip region was often superposed by the effects of speaking.

4.2. Emotion space classification

4.2.1. Classification into emotion space subclasses

The emotion space attributes were categorized into three distinct classes as described in Sec. 2.2.2. For *valence* the results obtained from using the eye region, lip region and their combination are given in Table 3 on page 8. The maximum average recognition rate of 80.1% was observed for the eye region, and the minimum of 72.0% was observed for the lip region. The combination of the two regions of interest gave an accuracy of 75.0% which lies between the two individual results.
For *activation*, Table 4 shows the classification results of the eye region, the lip region, and the combination of both. Once again the highest classification accuracy is obtained for the eye region, 80.1%, the lowest for the lip region, 70.4%, with their combination lying between the two results, 79.4%. For all cases the confusion of the extreme classes was with the neutral class rather than among themselves. A maximum of 83.6% classification accuracy for class *negative* of *valence* and a highest accuracy of 85.4% for class *excited* of *activation* was recorded.
In general the classification in the emotion space using three classes per emotion primitive yielded fairly good recognition rates. The results were better than the classification into emotion categories. This is in accordance with the human evaluation results (c.f. Sec. 2.2).

4.2.2. Continuous-valued emotion estimation

For a continuous valued estimate of *valence* and *activation* as a parallel to promising results in acoustic emotion recognition [15], a neuro-fuzzy emotion estimation method was implemented. For training, the continuous-valued references obtained from the mean estimate of the evaluators were fuzzified into three membership grades of the linguistic variables *negative, neutral, positive* for *valence*, and *calm, neutral, excited* for *activation*, respectively. The ANN was trained with these values. For testing, the ANN output neurons were regarded as membership grades of these linguistic variables. They were defuzzified using centroid method [24] to obtain a single crisp value.
For each image the error was calculated as a difference between the continuous-valued emotion primitives estimate and the evaluator reference, which both are in the range of [-1, +1]. The average error over all speakers is given in Table 5. The error was between 0.30 and 0.32 and thus approximately the same for both face regions as well as for their combination. These results are worse than those for emotion estimation using acoustic features [15]. They indicate that continuous-valued estimation of *valence* and *activation* from facial expressions is feasible but difficult.

Table 5. *Mean error of continuous-valued emotion primitives estimation.*

	Eye region	*Lip region*	*Eye and lip region combined*
Valence	0.32	0.31	0.31
Activation	0.32	0.30	0.30

5. META-FEATURES FOR EMOTION RECOGNITION IN FACIAL EXPRESSIONS

5.1. Motivation and method

In the case of our database, a large reduction of data took place due to the evaluator disagreement as well as the bad distribution of the emotions in the database into different classes. The poor evaluator agreement is probably due to the difficulty in classifying emotions into discrete classes. The analysis of emotion mixtures as primary and secondary emotions only marginally improved the agreement. This difficulty results from the spontaneous nature of the facial expressions, and from the fact that the images were taken from speakers in both the active and inactive speaking state. These conditions are reflected in the moderate recognition rates.

Thus as a next step we defined a number of emotionally relevant meta-features in the facial images which can more easily be determined by human evaluators. The meta-features defined for the eye region are shown in Fig. 8. They include *closed eyes* and *open eyes* as a coarse first-stage description, and *raised eyebrows, frowning,* and *not frowning* as a more detailed second-stage description. Since *raised eyebrows* in the case of *open eyes* was only present with very wide open eyes, we combined this appearance as a third, joint first-stage class *widely open eyes*. The meta-features defined for the lip region are shown in Fig. 9. They include *closed mouth, open mouth* and *tightly closed mouth* as a coarse first-stage description, and *smiling* and *not smiling* as a more detailed second-stage description.
Some of these meta-features are exact replica of the FACS-coded action units (AUs). For instance, *open eyes, frowning* is equivalent to AU4, and *raised eyebrows* is equivalent to AU2. The main reasons for choosing these meta-features are the apparent easiness of the evaluation and the expected relation to emotions expressed in the face.
Feature extraction was performed as described in Section 3. The features were passed to two cascaded ANNs for classification. The first one was to recognize the coarse first-stage class of the meta-feature, and the second classifier was to recognize the more

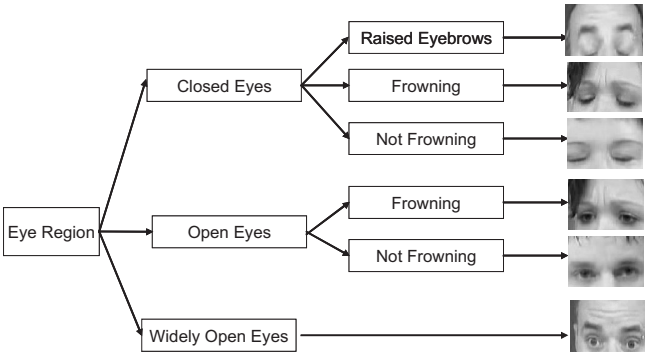

Fig. 8. *Meta-features chosen for the eye region.*

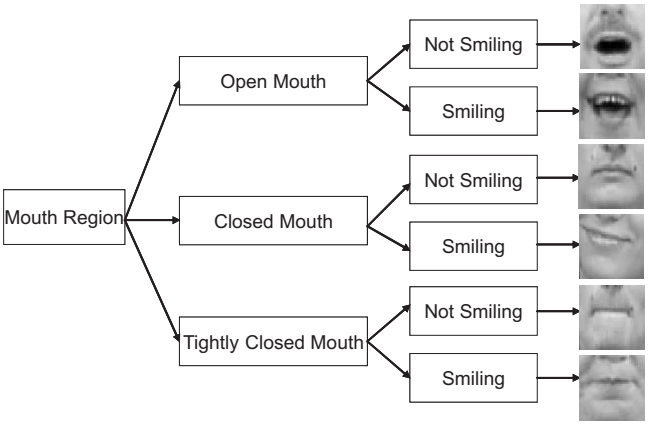

Fig. 9. *Meta-features chosen for the lip region.*

detailed second-stage class.

5.2. Data and results

For this feasibility study we have constructed a subset of 4 speakers of the VAM database. A minimum of 84 and 54 images, and a maximum of 150 and 180 images was available for the eye region and the lip region, respectively. The difference in the number of images is again to ensure that each class has the same number of images and also because a greater variety of pictures was not available in the VAM database. This dataset was evaluated by 4 evaluators with an evaluator agreement of 94.0%, which was significantly higher than for the evaluation tasks described in Sec. 2.2.

The results are reported in Tables 6, 7, and 8 for stage 1 and stage 2 classification, respectively, as average values over several speakers (see page 8). In addition to the recognition rates, the number of speakers and the number of images used is indicated, since these numbers varied for the different classification tests. From Table 6 it can be read that the average recognition rate for classifying the eye region into the stage 1-classes *closed eyes, open eyes,* and *widely open eyes* was 84.2%. Classifying the lip region into the stage 1-classes *closed mouth, open mouth,* and *tightly closed mouth* resulted in an average recognition rate of 81.7%, where *open mouth* was very well recognized (92.2%), and sometimes *closed mouth* and *tightly closed mouth* was confused.
For the stage-2 classifications, a smaller number of images (min 48, max 124) was available after equalization of the class distributions, but based on these images the recognition results were very good. *Frowning* vs. *not frowning* was correctly decided in 99.1% of the images with open eyes. The stage 2-classification of closed eyes into the classes *raised eyebrows, frowning*, and *not frowning (neutral)* was done with an average recognition rate of 89.6%. A comparable recognition rate was obtained when *open mouth* images were classified into *smiling* and *not smiling*, which was 92.7%. The classification accuracy of *closed mouth* images was 91.3% and thus marginally below the recognition rate for *open mouth* images. The subclassification of *tightly closed mouth* into *smiling* and *not smiling* is not shown as it was only available for one speaker in our database.
In general the recognition of these facial meta-features was significantly better than the recognition of emotions both in the emotion category and the emotion space approach. These results are in accordance with the human evaluation performance. Thus for spontaneous facial expressions of both speaking and non-speaking persons a 2-stage meta-feature classification followed by a mapping from these meta-feature to emotion classes seems to be the most promising approach.

6. DISCUSSION AND OUTLOOK

6.1. Discussion

In this paper we investigated the feasibility of emotion recognition in spontaneous facial expressions of both speaking and non-speaking persons. The emotion space method using the emotion primitives *valence* and *activation* was compared to the emotion categories method using the six basic emotions *anger, happiness, disgust, fear, sadness, surprise* and *neutral*. We have also explored the feasibility of continuous-valued emotion primitives estimation, which might be beneficial for emotion tracking, and for considering speaker dependent emotion biases. For feature extraction, multi-scale, multi-orientation Gabor filtering combined with a PCA was applied. Artificial Neural Networks, with a neuro-fuzzy wrapping for the continuous-valued estimation, was used for classification.
Both, the emotion categories and the emotion space classes could be recognized well above chance level. The maximum average recognition rate for emotion category classification was 72.9%. The maximum average recognition rate for emotion space classification was 80.1%. The mean error for continuous-valued emotion primitives estimation was 0.3, when the range of values was [-1,+1]. It was found that, for our data, the recognition rates were significantly below those reported in the literature, which were almost 100%.
The classification accuracy was better when the eye region features were used as input features. Classification accuracy was also better for the emotion subspace than for the emotion categories, though this is partly due to the smaller number of distinguished classes.
Much better results were obtained by using the meta-features described in Section 5, such as *open eyes* vs. *closed eyes* as a stage 1 meta-feature, and *open mouth, smiling* vs. *open mouth, not smiling* as a stage 2 meta feature. Using this 2-stage classification method, an average recognition rate between 81.7% and 99.1% was achieved for the individual classifications, which was much

closer to the results reported in the literature. Thus for spontaneous data as used in the VAM corpus, recognizing these facial meta-features is the most promising approach.

6.2. Outlook

In the future work, the meta-features should be mapped to emotion classes. Additionally, fusing of the facial and the acoustic emotion recognition should be done. The proposed methods should be verified on a larger dataset, and they should be compared to alternative algorithms, such as optical flow methods.

7. REFERENCES

[1] C. Darwin, *The Expression of the Emotions in Man and Animals*, Oxford University Press, New York, 1872, 3rd edition (1998).

[2] B. Fasel and J. Luettin, "Automatic Facial Expression Analysis: A Survey," in *Pattern Recognition*, 2003, pp. 259–275.

[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] C. Padgett and G. Cottrell, "Identifying Emotion in Static Face Images," in *Proc. of the 2nd Joint Symposium on Neural Computation*, La Jolla, CA, 1995, vol. 5, pp. 91–101.

[5] M.J. Lyons, J. Budynek, A. Plantey, and S. Akamatsu, "Classifying Facial Attributes Using a 2-D Gabor Wavelet and Discriminant Analysis," in *Proceedings of the 4th Int. Conf. on Automatic Face and Gesture Recognition*, Grenoble, 2000, pp. 202–207.

[6] M. Bartlett, G. Littlewort, I. Fasel, and J. Movellan, "Real time face detection and facial expression recognition: Development and application to human-computer interaction," in *CVPR Workshop on Computer Vision and Pattern Recognition for Human-Computer Interaction*, 2003.

[7] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding Facial Expressions with Gabor Wavelets," in *Proceedings of the 3rd Int. Conf. on Automatic Face and Gesture Recognition*, 1998, pp. 2000–2005.

[8] T. Kanade, J. Cohn, and Y. Tian, "Comprehensive Database for Facial Expression Analysis," in *Proceedings of the 4th Int. Conf. on Automatic Face and Gesture Recognition*, 2000, pp. 46–53.

[9] P. Ekman and W. Friesen, *Facial Action Coding System: A Technique for the Measurement of Facial Movement*, Consulting Psychologists Press, Palo Alto, 1978.

[10] G. Donato, M.S. Bartlett, J.C. Hager, P. Ekman, and T.J. Sejnowski, "Classifying Facial Actions," in *IEEE Trans. Pattern Analysis and Machine Intelligence*, 1999, pp. 974–989.

[11] Y. Tian, T. Kanade, and J. Cohn, "Recognizing Action Units for Facial Expression Analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 2, pp. 97–115, 2001.

[12] N. Sebe, M.S. Lew, I. Cohen, Y. Sun, T. Gevers, and T.S. Huang, "Authentic Facial Expression Analysis," in *Proceedings 6th Int. Conf. on Automatic Face and Gesture Recognition*, 2004, pp. 517–522.

[13] E. Douglas-Cowie, R. Cowie, and M. Schröder, "A new emotion database - considerations, sources and scope," in *Proc. ISCA ITRW on Speech and Emotion*, Newcastle, UK, 2000, pp. 39–44.

[14] S.V. Ioannou, A.T. Raouzaiou, V.A. Tzouvaras, T.P. Mailis, K.C. Karpouzis, and S.D. Kollias, "Emotion Recognition Through Facial Expression Analysis Based on a Neurofuzzy Network," *Neural Networks*, vol. 18, pp. 423–435, 2005.

[15] M. Grimm and K. Kroschel, "Rule-based emotion classification using acoustic features," in *Proc. 3rd International Conference on Telemedicine and Multimedia Communication*, Kajetany, Poland, October 2005.

[16] R. Kehrein, "The Prosody of Authentic Emotions," in *Proc. Speech Prosody Conf.*, 2002, pp. 423–426.

[17] M. Grimm, E. Mower, K. Kroschel, and S. Narayanan, "Combining Categorical And Primitives-Based Emotion Recognition," in *Proc. EUSIPCO*, Florence, Italy, 2006.

[18] M. Grimm and K. Kroschel, "Evaluation of Natural Emotions Using Self Assessment Manikins," in *Proceedings IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, San Juan, Puerto Rico, December 2005, pp. 381–385.

[19] P. Viola and M. Jones, "Robust Real-time Object Detection," in *Proc. of the 2nd IEEE Workshop on Statistical and Computational Theories of Vision*, 2001, pp. 1–25.

[20] K. Peng, L. Chen, S. Ruan, and G. Kukharev, "A Robust Algorithm for Eye Detection on Gray Intensity Face without Spectacles," *Journal of Computer Science and Technology*, vol. 5, pp. 127–132, 2005.

[21] Z. Zhang, M. Lyons, M. Schuster, and S. Akamatsu, "Comparison between geometry-based and gabor-wavelets-based facial expression recognition using multi-layer perceptron," in *Proceedings of the 3rd Int. Conf. on Automatic Face and Gesture Recognition*, 1998, pp. 454–459.

[22] J.J. Bazzo and M.V. Lamar, "Recognizing Facial Actions Using Gabor Wavelets with Neutral Face Average Difference," in *Proceedings of the 6th Int. Conf. on Automatic Face and Gesture Recognition*, 2004, pp. 505–510.

[23] T. Kanade, Y. Tian, and J.F Cohn, "Evaluation of Gabor-Wavelet-Based Facial Action Unit Recognition in Image sequence of increasing complexity," in *Proceedings of the 5th Int. Conf. on Automatic Face and Gesture Recognition*, 2002, pp. 229–234.

[24] K. Kroschel, *Statistische Informationstheorie*, Springer Verlag Berlin, 4. edition, 2004.

Table 2. *Recognition rates of emotion category classification.*

	Eye region						*Lip region*						*Eye and lip region combined*					
	A	H	D	Sa	Su	N	A	H	D	Sa	Su	N	A	H	D	Sa	Su	N
A	**71.5**	5.9	0	5.9	8.1	8.6	**62.9**	11.3	0	5.9	7.0	12.9	**69.3**	9.7	0	4.3	6.5	10.2
H	5.8	**70.9**	0.6	2.3	3.5	16.9	6.4	**77.3**	1.2	1.2	1.7	12.2	8.7	**70.9**	0.6	3.5	2.9	13.4
D	0	20.0	**70.0**	0	0	10.0	0	30.0	**40.0**	0	0	30.0	0	0	**50.0**	0	0	50.0
Sa	7.8	5.6	0	**76.7**	5.5	4.4	10.0	8.9	0	**63.3**	11.1	6.7	11.1	7.8	0	**67.8**	5.5	7.8
Su	7.2	6.4	0	5.4	**73.7**	7.3	13.6	7.3	0	4.5	**61.9**	12.7	6.4	3.6	0	5.4	**77.3**	7.3
N	10.7	10.7	1.0	1.5	1.5	**74.6**	21.9	20.4	2.0	5.6	9.2	**40.2**	8.7	12.2	1.0	1.0	3.1	**74.0**

Table 3. *Recognition rates of emotion space classification (valence).*

	Eye region			*Lip region*			*Eye and lip region combined*		
	Negative	Neutral	Positive	Negative	Neutral	Positive	Negative	Neutral	Positive
Negative	**83.6**	14.2	2.2	**69.4**	26.1	4.5	**74.3**	23.9	1.8
Neutral	12.4	**78.8**	8.8	19.8	**73.4**	6.8	14.4	**76.2**	9.4
Positive	5.0	17.0	**78.0**	7.1	19.8	**73.1**	7.1	18.4	**74.5**

Table 4. *Recognition rates of emotion space classification (activation).*

	Eye region			*Lip region*			*Eye and lip region combined*		
	Calm	Neutral	Excited	Calm	Neutral	Excited	Calm	Neutral	Excited
Calm	**79.3**	19.4	1.3	**71.5**	26.5	2.0	**82.6**	16.9	0.5
Neutral	18.6	**75.6**	5.8	26.0	**66.9**	7.1	16.4	**76.5**	7.1
Excited	4.6	10.0	**85.4**	4.6	22.5	**72.9**	5.4	15.5	**79.1**

Table 6. *Average recognition results for stage 1 classification of the eye region and the lip region, respectively.*

Eye region						*Lip Region*					
Recognition rate				No. of	No. of	Recognition rate				No. of	No. of
Closed	Open	Wide	Avg.	speakers	images	Closed	Open	Tight	Avg.	speakers	images
86.3%	84.3%	82.2%	84.2%	4	438	66.0%	92.2%	86.9%	81.7%	3	354

Table 7. *Average recognition results for stage 2 classification of the eye region classes open eyes and closed eyes, respectively.*

Open eyes					*Closed eyes*					
Recognition rate			No. of	No. of	Recognition rate				No. of	No. of
Frowning	Not frowning	Avg.	speakers	images	Frowning	Neutral	Raised eyebr.	Avg.	speakers	images
98.2%	100.0%	99.1%	4	112	81.3%	100.0%	87.5%	89.6%	2	48

Table 8. *Average recognition results for stage 2 classification of the lip region classes open mouth and closed mouth, respectively.*

Open mouth					*Closed mouth*				
Recognition rate			No. of	No. of	Recognition rate			No. of	No. of
Smiling	Not smiling	Avg.	speakers	images	Smiling	Not smiling	Avg.	speakers	images
93.5%	91.9%	92.7%	4	124	91.3%	91.3%	91.3%	4	94