EFFICIENT RECOGNITION OF AUTHENTIC DYNAMIC FACIAL EXPRESSIONS ON THE FEEDTUM DATABASE

Frank Wallhoff, Björn Schuller, Michael Hawellek, and Gerhard Rigoll

Technische Universität München, Institute for Human-Machine Interaction
Arcisstrasse 21, 80333 Munich

ABSTRACT

In order to allow for fast recognition of a user's affective state we discuss innovative holistic and self organizing approaches for efficient facial expression analysis. The feature set is thereby formed by global descriptors and MPEG based DCT coefficients. In view of subsequent classification we compare modelling by pseudo multi-dimensional Hidden Markov Models and Support Vector Machines. Within the latter case super-vectors are constructed based on Sequential Floating Search Methods. Extensive test-runs as a proof of concept are carried out on our publicly available FEEDTUM database consisting of elicited spontaneous emotions of 18 subjects within the MPEG-4 emotion-set plus added neutrality. Maximum recognition performance reaches the benchmark-rate gained by a human perception test with 20 test-persons and manifest the effectiveness of the introduced novel concepts.

1. INTRODUCTION

Affective computing has emerged as an independent research field within the Computer Vision society covering a wide range of applications [1, 2]. Therefore a broad variety of approaches has been presented. Most of them are lacking real time capabilities or are restricted to non-authentic caricatures only [3]. Although there is some diversity, Ekman et al. found evidence of inter-cultural universality of facial expressions, which can be categorised into happiness, sadness, anger, fear, surprise, and disgust [4]. This insight was further manifested by the fact of a MPEG-4 emotion-set .

For actual multimedia applications featuring emotion recognition based on video we therefore postulate the following constraints: First, a user state can be identified by observing the arising dynamics while changing from the neutral to the strongest appearance of an expression, the apex. Second, the expressions itself are natural and authentic. During a supervised training phase the complex patterns can be learned without time consuming labelling and transcription by using holistic and self organizing classifiers. Finally, for a system to run in real-time the overall computational efforts must be low.

These constraints opened the following road-map: Investigate a self learning statistical classifier which may even be complex in a first stage and is based on low computational features. Perform experiments on natural video material. Try to bridge the gaps between simple features, computational complexity and performance.

To fulfil the above constraints the pattern processing scheme is applied to a given video sequence as depicted in Figure 1.

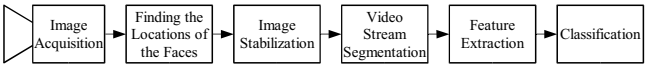

Fig. 1. Functional system breakdown

Independently of the kind of processed image material face locations are found in the first step. To avoid noise arisen from inprecisely located faces the found regions are additionally stabilized using block matching techniques. In a subsequent step the facial substreams are continuously scanned for expression boundaries and cut into segments to be classified without information about the semantic content. Hereafter the features of the pre-segmented face videos are extracted and led to a corresponding classifier.

To introduce the relevant subsystems the rest of this paper is structured as follows: In the first chapter the obeyed database and its characteristics are briefly introduced. Hereafter an algorithm to find segment boundaries independent of their content is presented. In the next section several common feature extraction techniques are recapitulated, followed by their corresponding classifiers. After the presentation of the experiments and results the paper concludes with a short discussion.

2. THE FEEDTUM EMOTION DATABASE

In the beginning of our first experiments no databases were freely available. Nowadays several non-commercial databases exist, but nearly all of them contain people acting emotions with unnatural caricatures. It is obvious that there are fundamental differences between acted sequences showing facial expressions and natural ones. Within the Face and Gesture network (FGNET) it was therefore decided to build up a publicly available database with mostly natural content for common use. Please find more information about the database and how to obtain it on-line at [5].

Besides the neutral state the content of the database covers the emotions anger, disgust, fear, happiness, sadness and surprise for each contained subject as shown in Figure 2, which have been recorded three times. Currently 18 subjects participated in the sessions.

Fig. 2. Examples of the 7 affective states in the database

To elicit the emotions as natural as possible it was decided to play several carefully selected stimuli videos and record the participants' reactions. For this purpose a video monitor together with a mounted camera on top were employed, which enables a direct frontal view. Both devices were controlled by a dedicated software that induced the desired emotions and started the recordings at the expected times.

In order to measure the quality of the acquired material and the difficulty to discriminate the basic emotions a perception test was performed with 20 subjects (half woman and half man) with ages between 23 and 38 years not being specialised to this task. It turned out that the right mean recognition performance was 61%, the worst 38% and the best 93%. This low performance can be explained by the fact that near natural facial expressions of unfamiliar persons without further context information has to be identified, which emphasises the challenge of the recognition task on this database.

3. VIDEO STREAM SEGMENTATION

For the latter classification in which it is likely that a certain emotional state appears video streams containing the facial region have to be pre-segmented. To assure low computational complexity which is mandatory for real-time applications the features and the decision rule used in this segmentation module have to be fast. In a similar scenario with the task to find action boundaries in meeting scenarios a common approach known within the speech recognition community, i.e. the Bayesian Information Criterion (BIC) [6], has already successfully been extended to the video domain. The fundamental idea is to compute if it is more likely that one model Φ_1 has generated an actual part of a video or that two models Φ_{21} and Φ_{22} changing at a certain time i were involved as depicted in Figure 3. The current segment with length n to be examined starts at frame s and ends at frame $s+n$. Each frame within this clip is represented by a feature e_t.

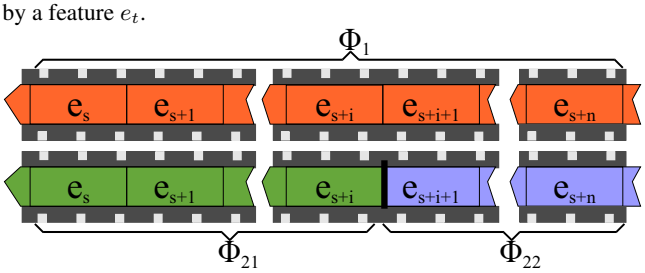

Fig. 3. Fundamental idea of the Video-Based Bayesian Information Criterion

Without knowing the semantic content a model Φ describing a clip can be derived by the covariance matrix Σ of seven dimensional vectors containing the centre of motion, the variance of the motion as well as the change over time in horizontal and vertical direction together with the global intensity of the change on difference images. The computation of these fast computable components is introduced in more detail later.

A semantic change within an examined part can be identified by finding the minimum of the ΔBIC yielding:

$$\Delta BIC_i = -\frac{n}{2}\log\|\Sigma_1\| + \frac{i}{2}\log\|\Sigma_{21}\| + \frac{n-i}{2}\log\|\Sigma_{22}\| + \frac{1}{2}\lambda\left(d + \frac{d(d+1)}{2}\right)\log n \quad (1)$$

The parameter λ represents an opportunity to control the sensitivity of the boundary detector. To process an entire video stream this process starts at the beginning with a minimum window size of 15 samples. If no change is detected the length is enlarged until a change has been detected. Then the process is repeated starting at the end of the predecessor until the video ends. For more detailed information please refer to [7].

4. FEATURE EXTRACTION

The quality of the extracted features representing the video content play a key role for the latter classification task. On the other hand a pattern recognition system using robust high level features is usually restricted to off-line applications. To enable a system achieving high recognition performance with affordable computational efforts we investigate two feature extraction approaches introduced below.

As a basis for both approaches it is assumed that an emotional state can be measured by the dynamics of the facial expressions which can effectively and holistically be measured using difference images. The active patterns in a video can be localized using sophisticated face detection and tracking algorithms [8]. To allow the presence of multiple faces the proposed techniques have to be applied in parallel queues independently. It is emphasized that beside the location of the active facial area no additional landmarks or segments have to be computed.

4.1. Global Motion

Video features are formed by a seven dimensional motion vector holistically describing the content as already introduced in the context of the automated scene boundary detection. Firstly difference images $I_d(x,y,t)$ between two subsequent images are computed and thresholded to remove noise and compression artefacts. Using this image motion features are then formed by the centre of mass, its change over time and the variance in both dimensions. Together with the global energy of the motion the feature vector can be formed: $\vec{x}_t = [m_x, m_y, \Delta m_x, \Delta m_y, \sigma_x, \sigma_y, i]^T$. Continuously repeating this procedure for a given video segment dynamic time series evolve with just a little computational effort. Typical examples and a formation of feature vectors over time are depicted in Figure 4. These sequences can be modelled in different ways as discussed later.

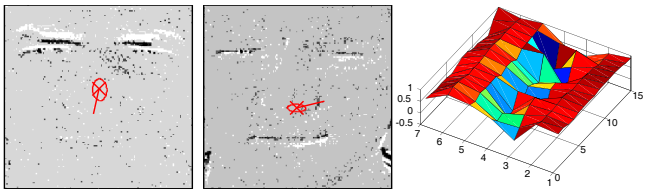

Fig. 4. Typical global motion vector sequence

4.2. Discrete Cosine Transform (DCT)

The background for introducing alternative features is motivated by the broadened availability of MPEG coded material. The time-invariance and linearity of the underlying (2D-)DCT $\mathcal{T}$ also enables a motion feature extraction applied to the available coefficients without additional computation by trivial differencing in the frequency domain directly:

$$\mathcal{T}_{\text{DCT}}\{I_d\} = \mathcal{T}_{\text{DCT}}\{I_1 - I_2\} = \mathcal{T}_{\text{DCT}}\{I_1\} - \mathcal{T}_{\text{DCT}}\{I_2\} \quad (2)$$

By the fact that the DCT is performed block-wise the spatial distribution of the motion can be preserved possibly containing additional relevant information. Motion blocks with overlap to other windows can also be derived efficiently in this domain [9].

For the latter modelling with so called Pseudo 3-dimensional Hidden Markov Models (P3DHMM) the resulting features for one image have to be scanned column-wise expanded by additional markers to indicate the start of a column. The entire feature pattern is formed by adding subsequent frames inserting extra start of image markers. A typical feature stream is shown in Figure 5.

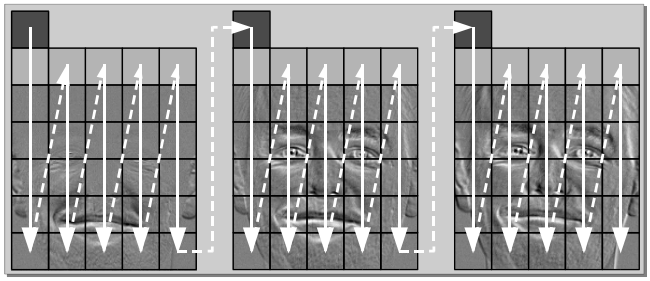

Fig. 5. Feature stream generation by element-wise differencing

4.3. Macro Motion Blocks

The third feature extraction combines attributes from above in the way that it can be computed rapidly while simultaneously preserving information about the spatial distribution of the motion. Therefore the area to be monitored is divided into several equally sized blocks which are processed using global motion features. Additional overlapping regions may be added. In Figure 6 a typical architecture consisting of 5 macro blocks is presented, where one motion frame consists of $5 \times 7 = 35$ elements.

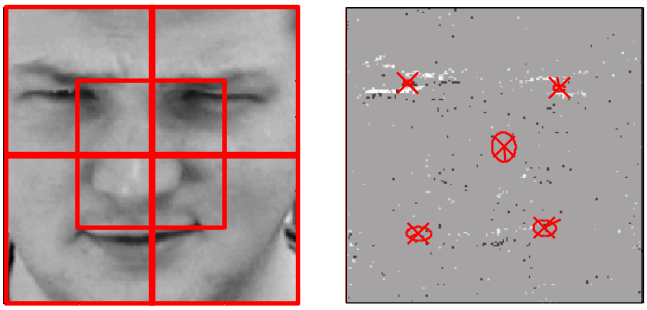

Fig. 6. Global motion macro blocks

5. CLASSIFICATION

Pattern classification is the last processing step in the functional system breakdown. In combination with the features introduced above two self learning approaches with supervised training phases are determined. The kind of the obeyed classifier heavily depends on the presented features, thus patterns with dynamic lengths can directly be classified with flexible Hidden Markov Models (HMM) and vectors with static properties may be processed with Support Vector Machines (SVM).

5.1. Hidden Markov Models

HMMs are well known especially within the speech recognition community but have also been reported to be superior in many other machine learning domains [10]. Due to their double stochastic modelling capabilities, i.e. warping over time and representing an observation exemplary by a normal distribution they can robustly represent the underlying structure of the data to be modelled. A HMM λ usually consists of a state transition vector a_{ij} and parameters describing the data distribution of the multivariate time series $\mathbf{X}$, e.g. means μ_i, variances σ_i and weights w_k of Gaussians. During the training phase these parameters are iteratively re-estimated according to the underlying data distribution to maximize the stochastic likelihood, which can be formulated as in Equation 3.

$$\lambda^* = \underset{\mathbf{X} \in \text{Training}}{\text{argmax}}\, p(\lambda(a_{ij}, \mu, \sigma, w)|\mathbf{X}) \tag{3}$$

An unknown pattern can be classified by computing the production probabilities for all known models from the emotion inventory choosing that with the highest production probability (Maximum Likelihood). For simple straight forward directed state transitions the HMM can be implemented very efficiently directly satisfying real time constraints in form of the well known Dynamic Time Warping algorithm, the state probabilities by Gaussian Mixture Models.

5.2. Pseudo 3-dimensional Hidden Markov Models

The so-called pseudo 3-dimensional Hidden Markov Models (P3D-HMM) can be interpreted as an extension of the linear models with respect to the state transition matrix and have already been successfully applied for gesture and emotion recognition problems [11].

Since real 3D models are computationally inexpensive the following restrictions have to be introduced. To model series of planar images every state of a linear HMM can be interpreted as a column of an image first, where the column itself it modelled by another column HMM. Such a pseudo 2-dimensional HMM can then be encapsulated into a third stochastic process to model the behaviour over time. Such models are also known as hierarchical HMMs. To assure a correct state alignment special marker-states as illustrated in Figure 5 have to be inserted denoted by dark and light shaded boxes.

Training and classification can be integrated analogous to the one-dimensional case. However, due to the dramatically extended number of states the classifier itself is expected to be rather time consuming.

5.3. Support Vector Machines

Besides HMMs diverse other machine-learning techniques are used in the field of emotion recognition [3, 12]. In [13] we made an extensive comparison including Naive Bayes, k-Nearest Neighbor classifier, SVM, Decision Trees, Artificial Neural Nets, and construction of ensembles as MultiBoosting or StackingC in the related field of speech emotion recognition. SVM have thereby proven the optimal choice. SVM are well known in the pattern recognition community and are highly popular due to their generalization capabilities achieved by structural risk minimization oriented training. Non-linear problems are solved by a transformation of the input feature vectors into a generally higher dimensional feature space by a mapping function, the *Kernel*, where linear separation is possible. Maximum discrimination is obtained by an optimal placement of the separation plane between the borders of two classes. The plane is spanned by *Support Vectors*. In general, SVM can handle only two-class problems. However, a variety of strategies exist for multi-class discrimination as couple-wise one-against-one decision, one-against-all classes or multi-layer decision. For more details refer to [14]. Herein, a polynomial kernel-function and couple-wise decision are used. For SVM classification a super-vector is constructed by concatenation of length normalised emotion patterns. To transform an feature sequence with dynamic length to a static one length normalization respective sub-sampling techniques are applied.

5.4. Feature Space Reduction

Having irrelevant features in the vector increases complexity for the classifier, and thereby directly decreases performance in most cases. This is especially true for sparse data, as the aimed at emotional data. It is therefore state-of-the-art to avoid this by reduction of the feature set by suited methods and could be shown highly effective in the

related field of speech emotion recognition [13]. Thereby also computational extraction effort may be spared as less features have to be calculated. Generally, feature reduction is done either by single feature relevance consideration, mostly done by filter-based selection as information gain calculation, or optimization of a feature set as a whole. Within the latter a classifier is used as optimization function, called wrapper, ideally the target one, and a search-function obligatory in most cases to ensure computability. As such wrapper-based search is superior in maximum obtained accuracy and efficiency we decided for Sequential-Floating-Forward-Search (SFFS) based feature selection [14] with use of the aimed at SVM (SVM-SFFS). SFFS is a Hill-Climbing search starting with an empty feature set adding the optimal next best feature within succeeding iterations. Floating search allows backward steps in order to avoid nesting effects.

6. EXPERIMENTS AND RESULTS

The above proposed feature extractions together with their corresponding classification paradigm are benchmarked on the FEEDTUM database using a 5-fold stratified cross-validation scheme. Table 1 provides an excerpt from representative results.

Feature	Classifier	Correct [%]
DCT	P3DHMM	33.61
Global Motion	1DHMM	38.66
Global Motion	SVM	42.33
Macro Motion Blocks	SVM	49.81
Macro Motion Blocks	SVM-SFFS	61.67

Table 1. Comparison of recognition performance

The multi-class SVM approach using reduced length normalized macro motion blocks achieves the best recognition performance with 61.67% which is slightly above the average of the user perception rate from 20 human subjects. Generally speaking the performance of the dynamic HMM based classifiers are worse compared to SVM, which might be explained by a too restricted size of the training corpus. Additionally the compromise between fast computable global and additional spatial information by using motion blocks is superior to all other features. By reducing the full feature set to 45 from originally 525 an additional gain was measured. Table 2 shows the confusion matrix of this setup. It turns out that the classes disgust, fear and sadness show the weakest performance, which is most probably caused by the rather low amplitudes of the facial motion between neutral an the apex.

Mean **61,67%**	*Anger*	*Disgust*	*Fear*	*Happiness*	*Surprise*	*Sadness*	*Neutral*	*Correct [%]*
Anger	**15**	4	1	1	0	3	2	57,7
Disgust	10	**12**	3	1	2	0	0	35,7
Fear	4	6	**11**	0	2	0	1	45,8
Happiness	1	0	1	**19**	1	1	3	70,3
Surprise	1	2	2	1	**17**	2	0	68,0
Sadness	3	2	0	2	0	**13**	5	52,0
Neutral	0	0	0	0	0	2	**24**	92,3

Table 2. SVM results after feature selection

The above observations indicate the following conclusion: Online recognition of natural facial expressions in video streams with fast computable macro block features in combination with SVM-SFFS leads to excellent results. Ongoing work is to expand the current emotion set with an additional *filler*-class to reduce the number of false assignments. Furthermore to allow for a more precise shape of the data to be modelled it is intended to add functionals obtained by static analysis of the global motion descriptors.

7. REFERENCES

[1] M. Stewart Bartlett, G. Littlewort, I. Fasel, and J. R. Movellan, "Real time recognition of facial expressions: Development and applications to human computer interaction.," *Computer Vision and Pattern Recognition*, 2003.

[2] P. Michel and R. El Kaliouby, "Real time facial expression recognition in video using support vector machines," in *Proc. of the 5th international conference on Multimodal interfaces*, Vancouver, British Columbia, Canada, 2003, pp. 258–264, ACM Press, 1-58113-621-8.

[3] M. Pantic and L. Rothkrantz, "Towards an affect-sensitive multimodal human-computer interaction," *Proc. of the IEEE*, vol. 91, no. 9, pp. 1370–1390, Sept. 2003.

[4] P. Ekman and K. Scherer (Editor), *Approaches to Emotion*, Lawrence Erlbaum Associates, London, 1984.

[5] F. Wallhoff, "The Facial Expressions and Emotions Database Homepage (FEEDTUM)," *www.mmk.ei.tum.de /~waf/fgnet/feedtum.html*, Sept. 2005.

[6] A. Tritschler and R. Gopinath, "Improved Speaker Segmentation and Segments Clustering Using the Bayesian Information Criterion," in *Proc. EUROSPEECH*, Paris, France, 1999, vol. 2, pp. 679–682.

[7] F. Wallhoff, M. Zobl, and G. Rigoll, "Action segmentation and recognition in meeting room scenarios," *Proceedings IEEE Intern. Conference on Image Processing (ICIP)*, Oct. 2004.

[8] F. Wallhoff, M. Zobl, G. Rigoll, and I. Potucek, "Face tracking in meeting room scenarios using omnidirectional views," *Proceedings Intern. Conference on Pattern Recognition (ICPR)*, Aug. 2004.

[9] Cherman L. Sabharwal and Brian Quandt, "An efficient algorithm for direct computation of adjacent block coefficients in the transformed domain," in *Proceedings of the 1997 ACM symposium on Applied computing*, San Jose, California, US, 1998, pp. 515–520.

[10] Lawrence R. Rabiner, "A tutorial on HMM and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.

[11] Stefan Müller, Frank Wallhoff, Frank Hülsken, and Gerhard Rigoll, "Facial Expression Recognition Using Pseudo 3-D Hidden Markov Models," in *16th Int. Conference on Pattern Recognition (ICPR)*, Quebec, Canada, Aug. 2002.

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

[13] B. Schuller, R. Jimenez Villar, R. Rigoll, and M. Lang, "Meta-classifiers in acoustic and linguistic feature fusion-based affect recognition," in *ICASSP*, Philadelphia, PA, USA, 2005, pp. 32–80.

[14] I. H. Witten and E. Frank, "Data mining, practical machine learning tools with java implementations," in *Morgan Kaufman*, San Francisco, 2000, p. 133.