SUBMOTIONS FOR HIDDEN MARKOV MODEL BASED DYNAMIC FACIAL ACTION RECOGNITION

Dejan Arsić, Joachim Schenk*, Björn Schuller, Frank Wallhoff and Gerhard Rigoll*

Technische Universität München, Institute for Human Machine Communication
Arcisstrasse 16, 80333 München, Germany
{arsic, schenk, schuller, wallhoff, rigoll}@tum.de

ABSTRACT

Video based analysis of a persons' mood or behavior is in general performed by interpreting various features observed on the body. Facial actions, such as speaking, yawning or laughing are considered as key features. Dynamic changes within the face can be modeled with the well known Hidden Markov Models (HMM). Unfortunately even within one class examples can show a high variance because of unknown start and end state or the length of a facial action. In this work we therefore perform a decomposition of those into so called submotions. These can be robustly recognized with HMMs, applying selected points in the face and their geometrical distances. Additionally the first and second derivation of the distances is included. A sequence of submotions is then interpreted with a dictionary and dynamic programming, as the order may be crucial. Analyzing the frequency of sequences shows the relevance of the submotions order. In an experimental section we show, that our novel submotion approach outperforms a standard HMM with the same set of features by nearly 30% absolute recognition rate.

Index Terms— Dynamic face expression recognition, gabor jets, HMMs, submotions

1. INTRODUCTION

On board security in aircraft cabins can be increased if we know how the passengers behave and what actions they perform during the flight. We aim to implement an automated video surveillance system, which sets out alerts, if unruly behaviors are detected. In [1] we have presented the functionality of such a system, which decomposes a complex behavior in several meaningful predefined indicators (PDI). With the help of psychologists and criminal experts in the project SAFEE we decided which PDIs are of major importance on board an aircraft. Among global motion, hand movement, and the use of tools, facial actions seem to be most significant. In contrast to other works we do not focus on face emotions as defined by Ekman [2], but on the activities laughing, speaking, yawning and other movements, for instance chewing or lip licking. Especially speaking and laughing may be detected better by audio, but in a noisy environment with a large number of people it is not possible to assign a sound to a single person.

In order to recognize PDIs we propose working on image sequences in which the position of facial features such as mouth, eyes, eyebrows and nose are tracked with Gabor Jets [3]. In a following step geometrical relationships between these features are computed, which results in a mesh over the face. Additionally the first and second derivation over time are determined.

In real world scenarios facial actions do not start with a fixed state and do not necessarily have constant transitions within the motion. The class of the facial activity may also change within as little as 3 frames. In order to cope with these constraints, we propose splitting up a facial activity in several so called submotions, similar to phonemes in speech recognition [4]. These are able to describe transitions between different facial states. For instance yawning may be described by opening the mouth, reaching a wide open mouth position, keeping it open and closing it slowly. Various different descriptions may be found for each activity. Therefore we suggest recognizing submotions with Hidden Markov Models, and a subsequent classification of the HMM output. We will present results of a frequency based approach and a distance measure, which considers the order of the appearance of submotions. Both approaches are able to compensate errors made during the submotion recognition tasks, and enhance the recognition results.

In this work we will present classification results with submotions on a behavior database, simulating events in an aircraft. Furthermore the recognition on complete actions without further decomposition has also been performed with HMMs, to evaluate the increase in performance after decomposition. The advantage of dynamic classification is shown by comparison to static classification with Support Vector Machines.

2. FACIAL FEATURES

In order to achieve a high precision and keep computation times as low as possible we decided to work with a small set of meaningful feature points in the face. This way we spare out a large part of the face, and reduce the required amount of data.

Based on the physiology of the face the MPEG-4 standard defines feature points, relevant for facial expression [5]. Out of the set of given points, 20, which may be detected automatically, have been chosen to describe faces. These are illustrated in figure 1 on the left side. Their coordinates are not generalizing faces in a person independent way. The size of eyes, mouth and eyebrows varies from person to person and the faces orientation is not considered. In a first step, the size of the faces is normalized and they are rotated into upright position. Therefore the angle between the eyes is computed, as they are usually on the same height. Afterwards the images are scaled to the same size, by aligning the distance between the both eyes.

Though the data is normalized there is still a large person dependency as well as variance between the different faces. Therefore we

*Both authors contributed to this work equally

This work has partially been funded by the European Union within the FP6 IST SAFEE Project. Special thanks to Thomas Mikschl for implementing the feature extraction.

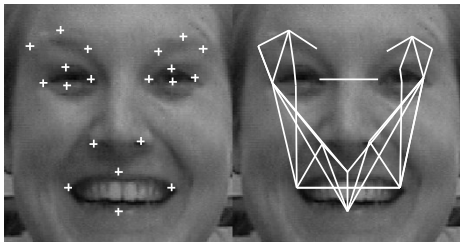

Fig. 1. Tracking of facial features (left) and the used geometrical description (right)

compute the distances of relevant points and create a mesh, which describes the geometrical relationship between the features. The 35 chosen distance measures are illustrated in figure 1 on the right hand side.
Considering that image sequences are available, it is reasonable to use information, retrieved out of motion within the face. These can be modeled either by the change of the labels' position or the distances between nodes in the mesh over the face. Speed and acceleration can be derived by computing the first and second derivation of the distances' differences over time.
As a result we can describe each face in a frame with a total of 125 features, which are: 20 $(x|y)$ coordinates, 35 distance measures, 35 speeds as well as 35 accelerations.

3. AUTOMATED FACIAL FEATURE TRACKING BY APPLYING GABOR JETS

Even though we use a face, eye and mouth detection based on a boosted set of Haar basis like features [6], precision is not sufficient for finding the exact position of the required feature points. Labeling the positions manually is not an option, as it is too time consuming and the goal should be an automated system.
The actual implementation requires a initialization of the 20 feature positions and is tracking the points near real time. Initialization is done manually at the moment, but can be automated in the future. For an already known feature position, Gabor Jets are computed with the Gabor kernel:

$$\psi_j(\vec{x}) = \frac{k_j^2}{\sigma_j^2} \exp\left(-\frac{k_j^2 x^2}{2\sigma^2}\right)\left[\exp\left(i\vec{k}_j\vec{x}\right) - \exp\left(-\frac{\sigma^2}{2}\right)\right] \quad (1)$$

Frequency and orientation of the filter are given by

$$\vec{k}_j = \begin{pmatrix} k_{jx} \\ k_{jy} \end{pmatrix} = \begin{pmatrix} k_\upsilon \cos\varphi_\upsilon \\ k_\upsilon \sin\varphi_\upsilon \end{pmatrix}, k_\upsilon = 2^{\frac{\upsilon+2}{2}}\pi, \varphi_\upsilon = \mu\frac{\pi}{8}. \quad (2)$$

With 5 frequencies and 8 orientations 40 jets are computed for each pixel.
In the subsequent frame tracking starts at the current position O. For this point and the surrounding ones the Gabor Jets are computed and compared with the Jet of the initialized point, by computing the Eucleadian distance. The one with the minimum distance is considered as the new position N.
Computation time can be reduced by limiting the size of the searching area. Therefore Jets of pixels in a concentric circle around the original point are compared with the initial one. If a minimum is found, the area around the actual minimum is examined for possible smaller differences, once more in a circle. Figure 2 illustrates this process.

This approach is independent of the actual orientation of the face, as long as the desired features are visible during the tracking process. Very fast movements in plane and depth are a major problem for this approach, as only small regions are observed, and a point, whose Gabor Jets have a minimum distance to the initialized one,is found anyway. Movements of the head in the image plane can be smoothened by performing a robust face detection in two subsequent frames. Though the faces are found robustly, the cut out is not necessarily the same. Afterwards in the area surrounding the face, a block matching algorithm is applied to compute the movement of the face [7].

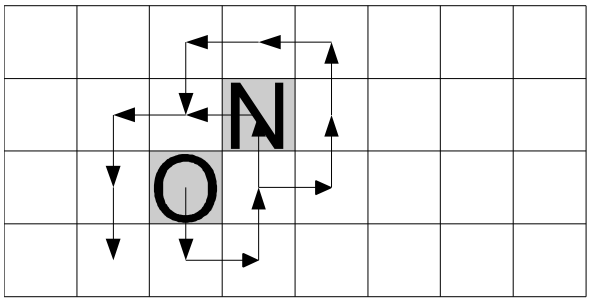

Fig. 2. Search of the subsequent feature position starting at O and ending at N

4. DATABASE

Common databases with facial behaviors, as listed in [8] usually contain the basic emotions suggested by Ekman [2]. Especially yawning and others, which may be any facial activity are not available in common databases. Consequently we decided creating a database with real facial actions. Natural laughing, speaking and other activities can be produced in an interview situation. However real yawning cannot be played, as most people consider them as a simple wide open mouth. Untrained persons can discriminate between a fake and a real one easily. Therefore some of the recordings were performed late at night or early in the morning, because of the high probability of having tired test subjects.

Fig. 3. Examples for speaking,laughing, yawning and other movements

During the recordings a flight situation has been simulated. We told the passengers what is happening on board, to capture behaviors. The test subjects where sitting in front of a camera, which covered face and upper body, and acted according to the announcements. Speaking can be considered real, also laughing, as the test subjects were slightly amused because of some of the announcements. The persons were also told not to suppress any actions, so we also could record some real yawns within the short sessions. After segmenting

the video material we chose sequences where the face was looking almost straight into the camera for at least 15 frames. Rotation was no issue. This way we were able to extract at least 101 samples for each of the classes, with an average length of 25 frames. The face is about 200×200 pixels large. Begin and end of the sequence were not fixed, so any transitions between facial activities may appear.
For each start frame we labeled the 20 relevant points manually and transcribed the submotions in the correct order.

5. INTRODUCING SUBMOTIONS

Fig. 4. Three different states within one yawn

Figure 5 shows three different stages of the facial action yawning. In this special selected example we can define a neutral start followed by a mouth opening and finally becoming a yawn. In a real world scenario it is not possible to segment captured data, to receive such precise start and end points and also the transitions between the states. As a result a short sequence of approximately one second may contain several different actions. For instance a laughing face may transform into a yawning one. If complete actions should be modeled, all this possible transitions have to be considered. With increasing number of facial activities, this leads to an exponential growth in required models.
Therefore we propose a decomposition of facial behaviors into so called submotions. These may contain transitions between different actions, the actions themselves and common characteristics. Obviously the number of required submotions is larger than the amount of facial actions we intend to detect. However further facial actions can now be added and simply described by the available submotions. This increase in actions no longer increases the required models. A major problem is the definition of submotions, which describe facial actions satisfactorily and are detectable. We collected a set of submotions manually, by inspecting video material. In order to show the performance of the decomposition, this should be sufficient. A entropy based automatic decision might be performed, if the submotion approach proves reliability.
Considering the limited amount of training data we decided to use only six submotions, whose order has been hand labeled within the training material. These were defined by the state of the mouth manually. In particular a closed, open, opening, closing, open mouth with mximum in horizontal / vertical direction were used. Start frame and end frame have not been assigned in the transcriptions. A sample of one second length can contain up to 5 submotions. Hence our database of 405 facial actions contains nearly 1500 submotions. Their length varies between 3 and 10 frames and is not depending on the actual class.

6. FACIAL ACTION RECOGNITION

The recognition task has been divided into two stages: Submotion recognition and the subsequent interpretation of those.

Submotion recognition

Hidden-Markov-Models (HMMs) [4] are applied for the first task, as these can cope with dynamic sequences with variable length, like our submotions. For training, we defined the order and number of submotions for each sample in our database. Neither the length, nor start and end frame are known to the training system. Each submotion is represented by a three or four state, left-right, continuous HMM and trained using the Baum-Welch-Algorithm [9]. During the training-process the submotions are aligned to the training data via the Viterbi-Algorithm in order to find the start and end frames of the contained submotions. For recognizing the sequence of occurring submotions, the Viterbi-Algorithm was applied again.

Facial action interpretation

In the second stage, we interpreted the sequence of submotions, gained form the previous stage to the facial actions. Therefore we developed a 'dictionary' containing facial actions and a number of possible representations in submotions. As the recognized submotion sequences suffer from insertion, deletion and confusion, a simple table lookup is not suited for the interpretation. Hence, we used dynamic programming to align the sequence of submotions in an optimal way to the corresponding facial actions. For instance a recognized sequence of submotions can be compared with 'dictionary' entries, by computing the Levenshtein Distance.
In this approach we assumed, that the order of submotions is crucial for the recognition of the represented facial units. To show that the frequency of the submotions within the sequences is not as significant as their order we used Support-Vector-Machines (SVMs) [10], which are a popular approach to solve two class problems. The vector contained the amount of appearances of each submotion and thus it had six features for our actual implementation. The feature-vector therefore represents the frequencies of submotions without an order. As shown in the results in Section 7 these frequencies are less significant for the interpretation of the facial actions. This proves that the facial action is determined rather by the order of the submotions than by their frequencies.

7. RESULTS

As a baseline system HMMs of the facial actions have been trained on sequences with various lengths. E. g. we derived a model for each of the facial units. Additionally a static approach with SVMs has been tested on whole facial actions. In this case we assumed that each facial action consists of the same amount of frames, as the number of input features has to be constant. All methods have been applied with four different feature sets, which contained combinations of distance measures (*d*), speed (*s*) and acceleration (*a*). Furthermore the impact of just the positions (*p*) has been tested.
Table 1 shows the results of a five folded cross evaluation. The upper half deals with recognizing facial actions as a whole – first by the SVM approach and second by using HMM (*HMM*) on our database. The lower half shows the recognition rate of the submotions (*SUB*) and their interpretation to facial actions via SVMs (*svm*) and dynamic programming (*dp*).
Within all approaches the distance measures seem to be more reliable than just the normalized coordinates, whereas the first derivation produces lower rates. Combining speed and distance will result in most reliable recognition. Adding acceleration has also a positive effect on the recognition rate, on the other hand the number of fea-

tures grows.
Comparing the classification of a whole sequence and the analysis of a series of submotions, the decomposition of facial actions performs far better than the standard approaches. A promising recognition rate of 90.1 % has been reached using all available features. With the sum set of features the HMM without decomposition only reaches 60.7 %. Hence our novel submotion decomposition leads to a recognition rate improvement of almost 30 % absolute. The results also show the importance of the interpretation of a sequence of submotion, as mistakes within the submotion recognition can be compensated. Although only 76.5 % of the submotions were recognized, the fusion yielded far better results.
The results also show the advantage of dynamic classification, as SVMs with a constant number of frames perform significantly worse compared to both HMM based approaches.

	p	*d*	*s*	*d+s*	*d+s+a*
SVM	51.5 %	63.7 %	35.0 %	68.3 %	59.2 %
HMM	58.6 %	64.2 %	44.3 %	67.8 %	**60.7** %
SUB	60.2 %	65.9 %	55.6 %	72.9 %	76.5 %
svm	69.9 %	79.6 %	42.8 %	82.3 %	83.6 %
dp	70.2 %	84.1 %	54.2 %	87.9 %	**90.1** %

Table 1. Overview on Recognition results. The best recognition rates have been yielded by recognizing submotions and aligning them via dynamic programming.

8. CONCLUSION AND OUTLOOK

In this work we presented a novel approach for the classification of facial actions. Faces are represented by 20 selected points, based on the MPEG-4 standard. From these we computed 35 geometrical differences between the points, creating a mesh over the face. Additionally the speed and acceleration of the change of the branches's length is computed.
The decomposition of facial actions into meaningful and detectable so called submotions, has been introduced. They represent different states of activities and transitions among them. Recognition is performed in two stages: First the classification of the submotions and then interpretation of their order or frequency. HMMs were trained for submotion classification with a combination of distance measures, speed, and acceleration. By applying a Viterbi-alignment the borders of the submotions within the training set were found, according to the given manually labeled transcription. The Recognition of the submotions of a test set was accomplished by a standard Viterbi-search, yielding recognition rates as high as 76.5 %.
In a subsequent step the sequence of detected submotions is interpreted to classify the corresponding facial actions. Both a frequency based matching and an alignment taking the order of the submotions into account have been tested. It turned out that a classification with SVMs, using the number of occurrences is clearly outperformed by dynamic programming, respecting the order of submotions. For this purpose submotion sequences were compared with 'dictionary' entries, by computing the Levenshtein distance. This way recognition rates up to 90.1 % have been reached. Compared to a holistic view of a sequence, with HMMs or even the analysis with a constant number of frames, this approach performed 30 % better.
In future work the database has to be extended both in size and number of classes, in order to distinguish between more actions. Additionally the submotions can be defined more precisely to generalize more possible facial behaviors. A possible solution could be an automated search within the given training set. At the moment continuous HMMs have been used. Quantization of the data and a transition to discrete HMMs may perform even better, which has still to be tested.
For a fully automated recognition a system for the robust initialization of the feature points has to be implemented. Subsequent to face detection these can be found applying Active Appearance Models [11].
The segmentation of a long video sequence, containing facial behaviors, by the presented system can then be used for a more complex behavior analysis. Analyzing longer sequences of both submotions and corresponding actions will result in a more precise description of a person's behavior. In quiet environments audio signals could be taken into account. Emotions can be classified very robust given a speech signal [12]. An audiovisual observation could enhance recognition.

9. REFERENCES

[1] D. Arsić, F. Wallhoff, B. Schuller, and G. Rigoll, "Video based online behavior detection using probabilistic multi-stream fusion," in *Proceedings IEEE International Conference on Image Processing (ICIP) 2005, Genoa*, Sept. 2005.

[2] Paul Ekman and Klaus Scherer, *Approaches To Emotion*, Lawrence Erlbaum Associates, 1984.

[3] Tai Sing Lee, "Image representation using 2d gabor wavelets," in *IEEE Transactions on Pattern Analysis And Machine Intelligence*, 1996, vol. 18.

[4] Lawrence Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," in *Proceedings of the IEEE*, 1989, vol. 77, pp. 257–286.

[5] J. Ostermann, "Animation of synthetic faces in mpeg-4," *Computer Animation*, pp. 49–51, 1998.

[6] P. Viola and M. Jones, "Robust real-time object detection," in *Second International Workshop On Statistical and Computional Theories of Vision - Modeling, Learning, Computing, and Sampling*, Vancouver, July 2001.

[7] D. Arsić, F. Wallhoff, B. Schuller, and G. Rigoll, "Bayesian network based multi stream fusion for automated online video surveillance," in *Proceedings EUROCON 2005, IEEE, Belgrade, Serbia and Montenegro*, Nov. 2005.

[8] http://emotion research.net/, Network of Excellence HUMAINE, 2004.

[9] L. E. Baum, "An inequality and associated maximalization technique in statistical estimation for probabilistic function of markov processes," in *Inequalities*, 1972, vol. 3, pp. 1–8.

[10] B. Schoelkopf, "Support vector learning," *Neural Information Processing Systems*, 2001.

[11] T. Cootes and C. Taylor, "Statistical models of appearance for computer vision," 2004.

[12] B. Schuller, S. Reiter, R. Mueller, M. Al-Hames, and G. Rigoll, "Speaker independent speech emotion recognition by ensemble classification," in *Proceedings 6th International Conference on Multimedia and Expo ICME 2005, Amsterdam*, 2005.