SUPPRESSION OF UNCERTAINTIES AT EMOTIONAL TRANSITIONS
Facial Mimics Recognition in Video with 3-D Model

Gerald Krell, Robert Niese, Ayoub Al-Hamadi and Bernd Michaelis
Otto von Guericke University Magdeburg, PO Box 4120, 39106 Magdeburg, Germany

Keywords: Man-machine communication, Associative deconvolution, Face analysis, Mimics, Emotion recognition.

Abstract: Facial expression is of increasing importance for man-machine communication. It is expected that future human computer interaction systems even include emotions of the user. In this work we present an associative approach based on a multi-channel deconvolution for processing of face expression data derived from video sequences supported by a 3-D facial model generated with stereo support. Photogrammetric techniques are applied to determine real world geometric measures and to create a feature vector. Standard classification is used to discriminate between a limited number of mimics, but often fails at transitions from one detected emotion state to another. The proposed associative approach reduces ambiguities at the transitions between different classified emotions. This way, typical patterns of facial expression change is considered.

1 INTRODUCTION

In recent years, there is a growing interest in human computer interaction systems. There is a tendency to develop technical systems which are able to adapt their functionality to the concrete user and therefore to assist the human much better than conventional machines. So called ’companions’ should help people to solve a certain task.

Such companion systems must be able to communicate with the operating user on a high level. The communication should take place as natural as possible. Therefore, we focus on human communication on the basis of speech, gesture and mimics. We call such systems which are able to recognize information on this level as cognitive systems.

Besides gestures and speech one important part of man-machine interaction which is addressed by the desired companion is the recognition of facial expressions of the human operator. Facial expression is one data mode to derive emotional information. It indexes physiological as well as psychological functioning. For instance, in previous works it has been used for the monitoring of the patient state after a surgery: The awaking phase and possible pain is detected in the recovery room (Niese et al., 2009) .

For reliable facial expression detection, the measurement method must be accurate and robust in order to derive suitable features. Also the temporal resolution must be high enough to detect changes of facial expression. These requirements lead to a video camera-based recognition system. We use a 3-D face model from stereo for pose estimation and determination of features to establish a classification basis for the facial expression recognition. Furthermore, the facial expression detection should take place under natural conditions. This demand prohibits attachment of markers to the facial skin. The observed person shouldn’t be restricted in his movement too much; head position can therefore vary pretty much. Also illumination should be normal - the surface detection can therefore not be based on structured light. These side conditions represent additional challenges for the measurement system.

The classified facial expressions are the basis to estimate the emotion of the person. The correlation of facial expression with the expected emotion can be verified by other biological measures (see (Ekman, 1994) for example).

In the paper, we investigate color image sequences of facial expressions which enables us to observe their dynamics. It is shown that typical temporal dependencies of facial expressions exist which therefore should be considered in order to increase the robustness of the correct recognition.

Krell G., Niese R., Al-Hamadi A. and Michaelis B. (2010).
SUPPRESSION OF UNCERTAINTIES AT EMOTIONAL TRANSITIONS - Facial Mimics Recognition in Video with 3-D Model.
In *Proceedings of the International Conference on Computer Vision Theory and Applications*, pages 537-542
DOI: 10.5220/0002852805370542

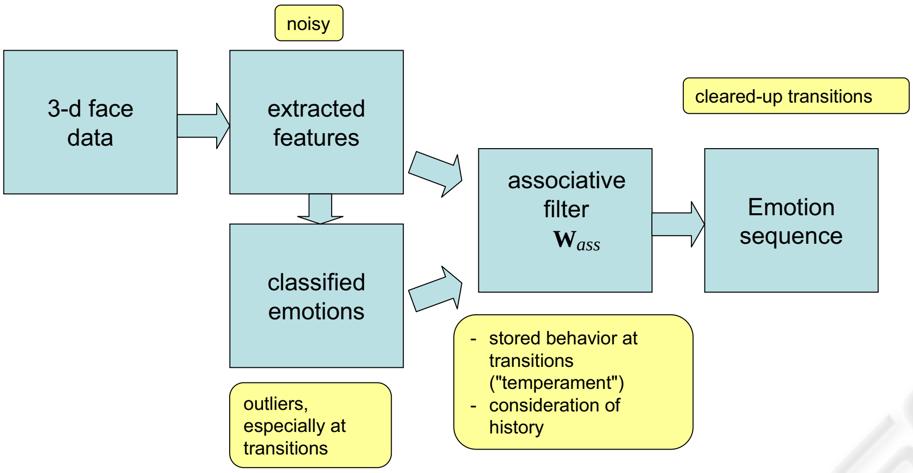
noisy
3-d face data
extracted features
classified emotions
outliers, especially at transitions
associative filter $\mathbf{W}_{ass}$
cleared-up transitions
Emotion sequence
- stored behavior at transitions ("temperament")
- consideration of history

Figure 1: Overview of mimic recognition with associative temporal processing.

2 RELATED WORK

Digital analysis of faces in images has received substantial effort in recent years. Applications include face recognition, facial expression analysis, face simulation, animation etc. (Li and Jain, 2001). Facial expressions are the natural way to express and to recognize emotions in human communication. Extensive studies laid a strong basis for the definition of universal facial expressions. Approaches to facial expression analysis using still images and video sequences usually apply a common hierarchy which is firstly to extract features from the facial images. Please refer to (Niese et al., 2008) for a more detailed literature survey on facial feature extraction.

The processing step of feature extraction is usually followed by the classification module. State of the art techniques for facial expression classification are often limited to more or less frontal poses and small skin color variations. Further, often they are sensitive to global motion with resulting perspective distortions. These restrictions have impact on the applicability, i.e. for human machine interfaces.

In the used method for feature extraction these restrictions are avoided by inclusion of photogrammetric measuring techniques and feature normalization. Most works focus on the recognition of facial expressions in still images or independently in frames of an image sequence.

As an alternative to this approach, in this paper processing of the classification probabilities is shown. By including the transitions from one facial expression to another, temporal dependencies are included in the evaluation of the feature data.

Typical changes can be considered as something like the emotionality which is directly linked with the ”temperament” of a particular person.

(Fragopanagos and Taylor, 2005) uses a neural network approach for the fusion of different data modalities in emotion recognition. We apply a multichannel deconvolution method for the estimation of the association filter in order to consider dependencies in the time series of emotional features and classification probabilities.

3 COMBINING CLASSIFICATION AND TEMPORAL PROCESSING

In this paper, a video-based system for the recognition of facial expressions with temporal processing of the classfication probabilities is used. The sketch of the overall system is shown in Fig. 1.

Additionally to the color image sequences 3-D context information is used. In particular, photogrammetric techniques are applied for the determination of features provided by real world measures. In this way, we achieve independence of the current head position, orientation and varying face sizes due to head movements and perspective foreshortening like in the case of out of plane rotations.

The recognition is based on the detection of facial points in the image. A camera and surface model is applied to determine the face position and establish transformations between 3-D world and image data. On the basis of detected facial feature points a normalized feature vector is built and fed to a SVM classifier (Chang and Lin., 2009).

The proposed representation of facial feature data leads to a superior classification of different expressions under real world conditions compared to other approaches (Cohen et al., 2003). The feature extraction and classification part of the system is described more in detail in (Niese et al., 2008) but the main ideas are also shown here.

The number of expressions to detect is dependent on the application but is strongly affected by the available training data. Here we discriminate five classes as a typical case in face analysis. Additional to four basic facial expressions we define the neutral expression.

Due to measuring errors the extracted features are of limited accuracy and can therefore be considered as 'noisy'. The classified emotions possess outliers, especially at the transition from one recognized emotion to another. The resulting feature vector and classified emotions are therefore passed to a post-processing system to include a-priori knowledge on the dynamics of emotions. This system could be considered as an associative memory which is able to restore incomplete or noisy data. We show a realization with a multi-channel-deconvolution method processing the time series of emotion classification probabilities of the different channels.

3.1 Training and Test Data

We trained the SVM classifier on our database, which contains about 3500 training samples (images) for the facial expressions of ten persons. Training has been done for five facial expressions including the neutral face.

The measured persons had the task to imitate a given emotion which is referred to as the expected emotion below. It can be represented by 5 temporal vectors with binary elements for each time step.

The shown classification results are based on 1000 test samples from 20 image sequences of about 50 frames length, starting from the neutral expression.

We use data of different persons than in the training phase for the test. The test scenarios contain position variations including out-of-plane rotation.

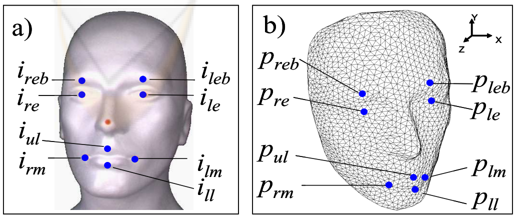

Figure 2: Facial feature points, a) Image, b) World projection.

3.2 Feature Extraction

The facial expression recognition is based on fiducial points, which are used to establish correspondence between model and real world data are defined and referred to as model anchor points (Fig. 2).

The initial Adaboost based face detection (Viola and Jones, 2001) is used to restrict the facial feature points search space. Application of the face detector is only required for re-initialization of the system.

The search space is limited by the previous face position and the extraction of the eye center points $[i_{le}, i_{re}]$ and mouth corner points $[i_{lm}, i_{rm}]$ follows. In the 2-D image processing part of the system, additionally two eyebrow points $[i_{leb}, i_{reb}]$ are detected.

The eye, eye brow and mouth detection is based on evaluation of color information and application of morphological operators. Assumptions regarding face geometry are included and image enhancement is applied.

This method has shown to also work under changing conditions, like in the appearance of teeth. Mouth and eye brows are detected reliably.

In order to transform image points to real world coordinates a surface model is defined which is based on an initial registration step where the face is captured once in frontal pose with neutral expression. Depth values are retrieved by measuring the distance from the camera plane to the surface model at the present world pose. A person specific mesh surface model of the face is established in this way which consists of a set of vertices and triangle indices (Fig. 2b).

In this face model, the nose tip can be found as the extreme point in front of the face. This 3-D point is then projected back and tracked in the 2-D image sequences. Together with the two eye positions the head position can be determined uniquely by this way.

Based on the point set we determine the ten-dimensional feature vector. The features comprise six Euclidean 3-d distances across the face (eye - eye brow - mouth for both sides, mouth width and height) and four angles, which are used to describe the current mouth shape. The transformation into world coordinates of a face model is schematically shown in Fig. 2b: the feature points found in the (2-D) images $[i_{le}, i_{re}, i_{lm}, i_{rm}, i_{leb}, i_{reb}]$ are used to update the corresponding 3-D model points $[p_{le}, p_{re}, p_{lm}, p_{rm}, p_{leb}, p_{reb}]$.

In the initial registration step of our system, we capture the face of the subject with neutral expression. After the surface model has been created, the neutral feature vector is determined. On the basis of this neutral configuration the current expression is analyzed.

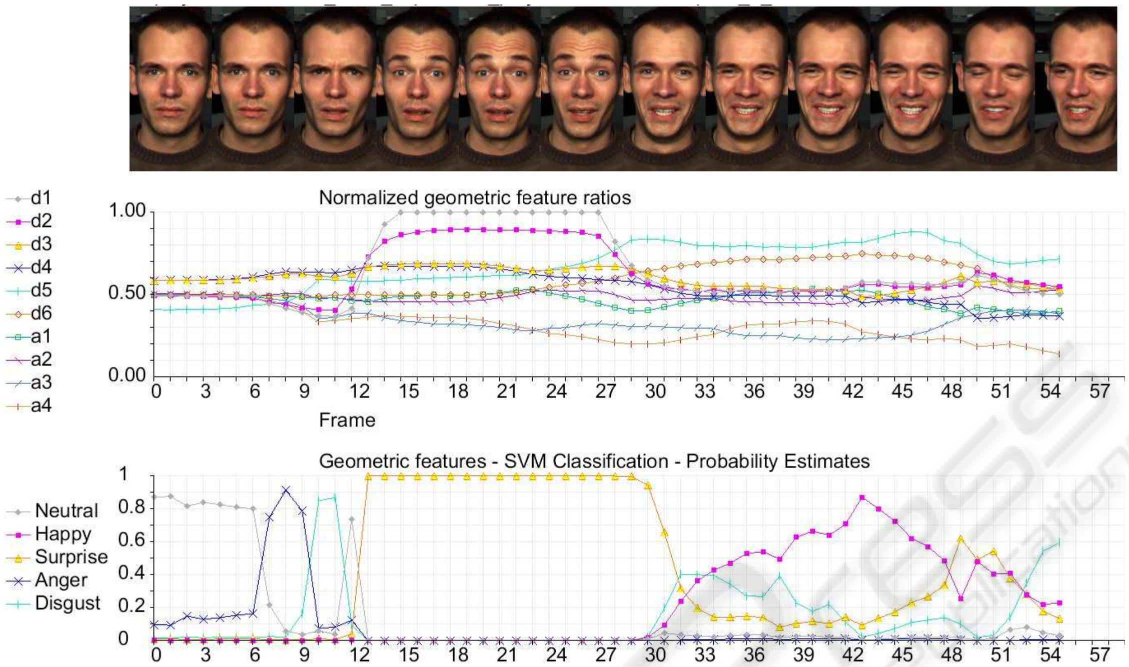

Figure 3: Example sequence: ’Neutral’-’Surprise’-’Happy’.

3.3 Classification

After building the feature vector, classification of the expression is performed. We assumed a number C of fundamental facial expressions. For the training and classification, a Support Vector Machine algorithm of the package LIBSVM has been applied (Chang and Lin., 2009) to the normalized feature data.

The SVM returns the class for each input sample. We further compute the probability for every class based on the training model and the method of pair wise coupling (Wu and Lin, 2004) giving the input for the post-processing unit.

3.4 Temporal Emotion Processing

In order to take temporal effects of emotion changes into consideration we developed a deconvolution method which is applied to the multi-channel system of acquired emotions. Deconvolution techniques are well known in signal and image processing (e.g. (Jain, 1998)). We developed the idea for the multi-channel system of emotions.

According to the test set-up (see 3.1) we can define the expected emotion vector

$$\mathbf{e}_{\mathrm{c}}^{\exp} = \left[\begin{array}{ccccc} e_{c,1}^{\exp} & \cdots & e_{c,i}^{\exp} & \cdots & e_{c,N}^{\exp} \end{array} \right]^{t}$$

containing the time-series of emotion values of channel c the test person has to imitate at time step i. This way a training target of the temporal processing system is given (ground truth).

t denotes the transposed. It has N elements according to the number of time steps (number of pictures in the training sequence).

In our case, we assumed bi-level emotions and a single emotion at a certain time (exclusive emotions). This is, the dicrete time function in the vectors $\mathbf{e}_{\mathrm{c}}^{\exp}$ are binary steps and exactly one of the channels has the value 1 at a certain time index i whereas all other channels are zero. This corresponds to the idea that the test person has to imitate one single facial expression at a certain time

$\mathbf{e}_{\mathrm{c}}$ is the corresponding vector of classification probability values of emotion channel c which has been obtained by measuring, feature extraction and classification as described above. Optimal multichannel FIR Filters $\mathbf{w}_{c,1\cdots C}$ for every emotion channel are estimated by the optimal (least mean square) solution of the equation system:

$$\mathbf{e}_{c}^{\exp} = \sum_{q=1}^{C} \left(\mathbf{e}_{q} * \mathbf{w}_{c,q} \right) + \mathbf{r}_{c} \qquad (1)$$

with $*$ the discrete convolution operator. $\mathbf{r}_c$ is the residual vector in channel c that is the approximation error when solving the over-determined equation system of equation (1)).

The weight vectors $\mathbf{w}_{c,q} = \left[\begin{array}{ccccc} w_{1}^{c,q} & \cdots & w_{j}^{c,q} & \cdots & w_{M}^{c,q} \end{array} \right]^{t}$ of length M according to the number of FIR filter taps consist of

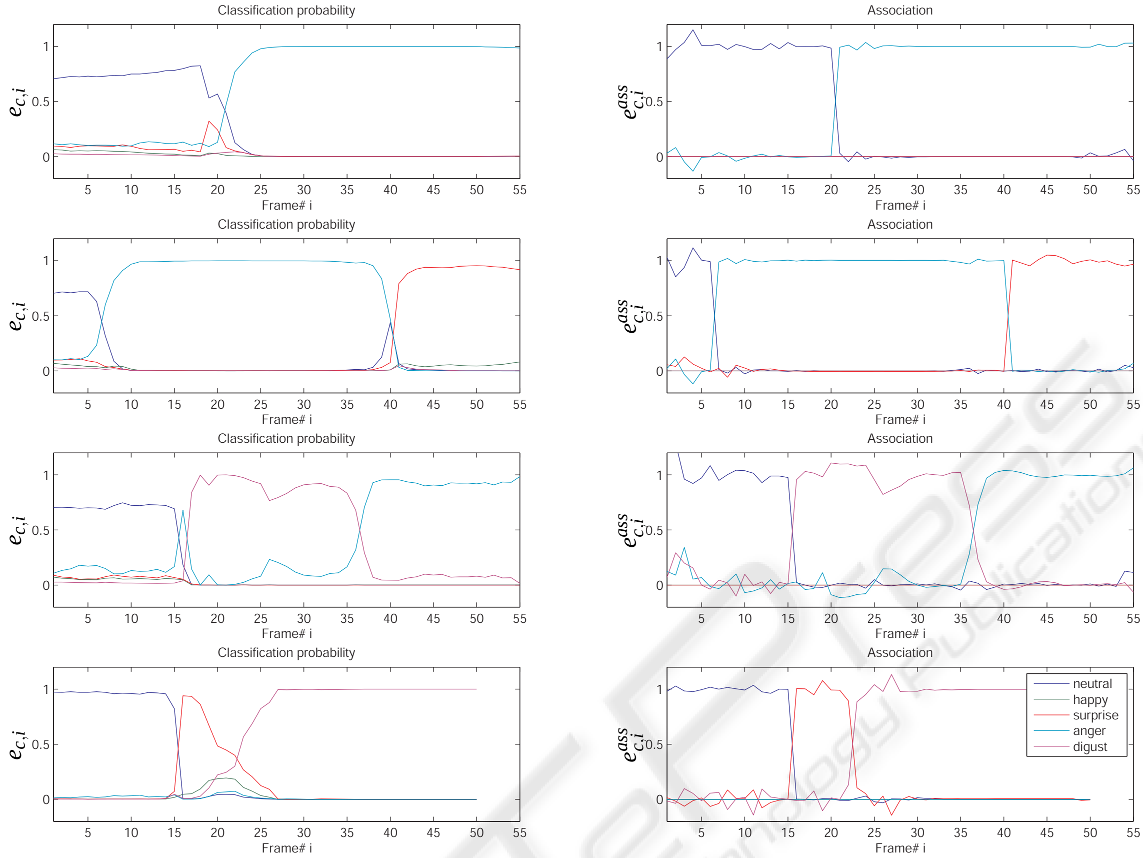

Figure 4: Classification probabilities for 4 selected sequences: original values (left), association (right).

the filter coefficients $w_j^{c,q}$ with index j and create the multi-channel filter matrix

$$\mathbf{W}_{ass} = \begin{bmatrix} \mathbf{w}^t_{1,1} & & & & \mathbf{w}^t_{1,C} \\ & \ddots & & \iddots & \\ & & \mathbf{w}^t_{m,n} & & \\ & \iddots & & \ddots & \\ \mathbf{w}^t_{C,1} & & & & \mathbf{w}^t_{C,C} \end{bmatrix} \qquad (2)$$

which thus is of size $M \cdot C$ by C. The calculated $\mathbf{W}_{ass}$ can be considered as an associative memory because it stores the associations for typical emotional transitions and how they can be interpreted. It is able to suppress false classification results when the recognized facial expression changes. The filters with index $m = n$ are the direct couplings whereas the other filters are responsible for the cross-coupling between the channels. This way the emotion channels are weighted automatically according to the expected emotion.

The result of temporally processing the classification probabilities is obtained by recalling the emotion associations:

$$\mathbf{e}^{ass}_c = \sum_{q=1}^{C} \left(\mathbf{e}_q * \mathbf{w}_{c,q}\right) \quad . \qquad (3)$$

4 EXPERIMENTAL RESULTS

In our problem we assumed five fundamental expressions: ’Happy’, ’Surprise’, ’Anger’ and ’Disgust’ and the neutral expression ’Neutral’ establishing our emotion channels ($c = 1 \cdots C, C = 5$). These results also include rotation of head and back and forth movement scenarios. The confusion matrix of the classification part (Niese et al., 2008) shows high recognition rates of over 90 percent for each class, which is presented by diagonal elements, but also a certain mixing between different classes expressed by the non-diagonal elements, in particular at the time of transition from one emotion to another.

In order to further improve the classification result we applied multi-channel deconvolution to the probability values. In the example, acausal association filters for each channel have been estimated that convolve a temporal neighborhood of 7 emotion samples ($M = 7$). For the consideration of multi-channel effects each of the C filters has therefore $C \cdot M = 35$ taps.

Fig. 3 shows a sample sequence. The image sequence (above) starts with the neutral facial expression. The candidate has then to imitate suprise and happiness. Normalized feature data is shown in the middle providing the input for the SVM classifier.

Below the classification probability values estimated by pairwise coupling corresponding to the detected features are depicted. It is clearly shown that at the constant parts of the facial expression time line the classification gives stable results. But at the transition from neutral to surprise false classifications are visible (Anger-Disgust-Neutral). When returning from the happy to the neutral expression surprise and disgust are detected which not has been intended.

Fig. 4 (left) shows the application of the associative deconvolution to the classification probability values. The results show again how the SVM is clearly defining boundaries between different classes when the facial expression is constant but fails at the transitions. All time series start with the neutral expression. In the upper sequence the expression changes to ’Happy’. In the second example two transitions are included: from ’Neutral’ to ’Happy’ and from ’Happy’ to ’Surprise’. The third example shows the transitions ’Neutral’-’Disgust’-’Anger’ and the forth ’Neutral’-’Surprise’-’Disgust’. In the right half of Fig. 3, the association of the probabilities values is shown. It is obvious that false classifications at the transitions from one emotion to the other are well suppressed.

5 CONCLUSIONS AND OUTLOOK

We have presented an efficient framework for facial expression recognition in human computer interaction systems. Our system achieves robust feature detection and expression classification and can also cope with variable head poses causing perspective foreshortening and changing face size of different skin colors.

The shown approach with a linear multi-channel deconvolution shows the principle of inclusion of temporal behavior. Additional inclusion of feature data together with the classification data should further improve the results. Current work is also attempting to estimate additional dynamic features. Those features are obtained by methods of motion analysis, e.g. optical flow techniques.

It is expected that other (non-linear) approaches such as associative memories, known from the artificial neural networks [e.g. (Kohonen, 1995)], could be interesting.

ACKNOWLEDGEMENTS

This work has been supported by DFG-Project TRR 62/1-2009.

REFERENCES

Chang, C.-C. and Lin., C.-J. (2009). Libsvm: a library for support vector machines.

Cohen, I., Sebe, N., Garg, A., Chen, L., and Huang, T. (2003). Facial expression recognition from video sequences: Temporal and static modeling. *Computer Vision and Image Understanding*, 91(1-2):160–187.

Ekman, P. (1994). Strong evidence for universals in facial expressions: a reply to russell’s mistaken critique. *Psychol. Bull Journal*, pages 268–287.

Fragopanagos, N. and Taylor, J. G. (2005). Emotion recognition in human-computer interaction. *Neural Networks, Special Issue*, 18:389–405.

Jain, A. (1998). *Fundamentals of Digital Image Processing. Prentice Hall*. Prentice Hall.

Kohonen, T. (1995). *Self-Organizing Maps*. Springer.

Li, S. and Jain, A. (2001). *Handbook of Face Recognition*. Springer.

Niese, R., Al-Hamadi, A., Aziz, F., and Michaelis, B. (2008). Robust facial expression recognition based on 3-d supported feature extraction and svm classification. In *Proc. of the IEEE International Conference on Automatic Face and Gesture Recognition (FG2008, Sept. 17-19)*.

Niese, R., Al-Hamadi, A., Panning, A., Brammen, D., Ebmeyer, U., and Michaelis, B. (2009). Towards pain recognition in post-operative phases using 3d-based features from video and support vector machines. *International Journal of Digital Content Technology and its Applications*, (in print).

Viola, P. and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition*.

Wu, T. and Lin, C. (2004). Probability estimates for multi-class classification by pair wise coupling. *Journal of Machine Learning Research*, 5:975–1005.