Adaptation of Facial Feature Extraction and Rule Generation in Emotion-Analysis Systems

S. Ioannou, A. Raouzaiou, K. Karpouzis and S. Kollias
Department of Electrical and Computer Engineering
National Technical University of Athens,
9, Heroon Polytechniou street, 157 80 Zographou, Greece
Phone: +30-210-7723039, Fax: +30-210-7722492
email: {sivann, araouz, kkarpou}@image.ntua.gr, stefanos@cs.ntua.gr

***Abstract*—The paper addresses the problem of emotion recognition in faces through an intelligent neuro-fuzzy system, where the extraction of facial features follows the MPEG-4 standard and is adapted to particular environmental conditions and specific persons. These features are associated to symbolic fuzzy predicates providing the classification of facial images according to the underlying emotional states. For this classification we use rules extracted from psychological studies and expression databases including extreme expressions such as those illustrated in Ekman's database. The rules are then refined in realistic conditions, taking into account the extracted features. The experimental results, based both in extreme and naturalistic databases developed in the frameworks of IST ERMIS and NoE HUMAINE, illustrate the capability of the developed system to analyse and recognise facial expressions in human computer interaction applications.**

I. INTRODUCTION

In this paper we develop a prototype system for human computer interaction that can interpret its users' attitude or emotional state, e.g., activation/interest, boredom, and anger. Visual features are extracted to guide this decision, supported by linguistic and paralinguistic analysis. For humans, facial gestures are very indicative for a persons' emotional state, so it's clear this extra information could have a serious impact on the overall performance of the system. However, extracting meaningful features is not straightforward.

The main goal is to extract some meaningful features from the face (the MPEG-4 facial animation and facial definition parameters), after having automatically detected the face in the images. Once the features extracted, we attempt to recognise the basic emotional states.

The aim of this work is to develop a system that works in realtime, i.e., with at most a few seconds delay between capturing the video data and generating the output. For the time being all processes work offline, in batch-mode, processing individual images or a whole video sequence at once and only then generating the output.

Test video material has been distributed of two actresses and shows a variety of expressions and emotions.

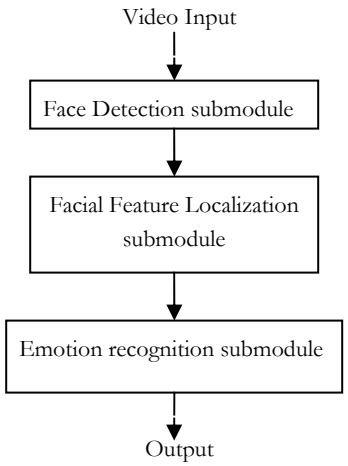


Figure 1. Overview of the facial gesture analysis and feature extraction module

II. FACE DETECTION

A. Technical description of the submodule

Given an arbitrary image, the goal of face detection is to determine whether or not there are any faces in the image, and if present, return the image location and face extent.

In this work we have used a face detection technique [1] based on support vector machines. Face detection module is used as initialization and, to recover from errors, is repeated e.g., every second. Its output is fed into the facial feature extraction module as can be seen in Figure 1.

B. Experimental Results

Figure 2 shows the result of running the face detection submodule on a frame of the data set created in the framework of IST Project ERMIS. The images (b), (e) indicate which pixels were classified as skin color (pixels shown in black). Based on this skin color detection, as well as on a variance detector, it's possible to reject about 90 % of the windows before any heavy computation starts. This makes the system perform at reasonable speed, by focussing on interesting parts in the image.

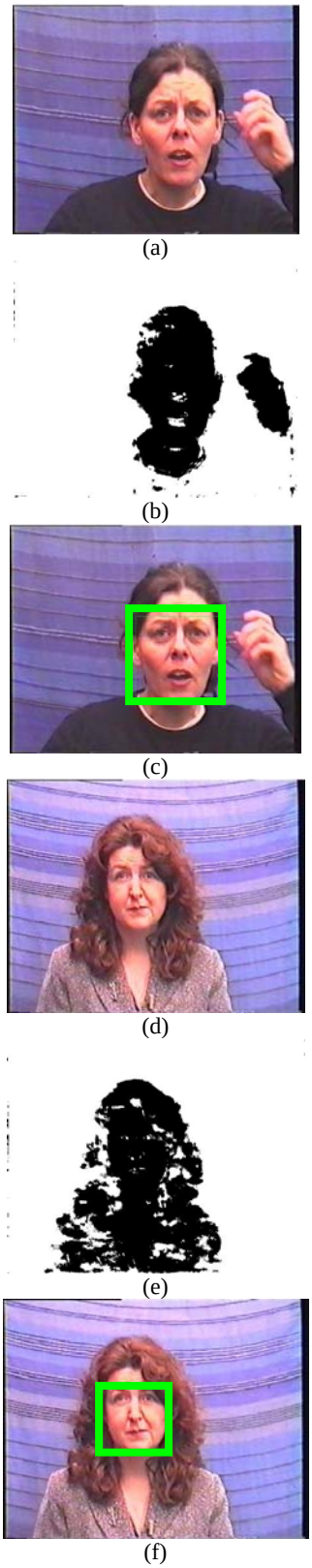


Figure 2. Face detection example ((a), (d): original image, (b), (e): skin color detection, (c), (f): detected face)

III. EMOTION ANALYSIS SYSTEM

A. FAP and FDP localization

1) Facial Features Relevant to Expression Analysis

Robust and accurate facial analysis and feature extraction has always been a complex problem that has been dealt with by posing presumptions or restrictions with respect to facial rotation and orientation, occlusion, lighting conditions and scaling.

In the framework of MPEG-4 standard, parameters have been specified for Face and Body Animation (FBA) by defining specific Face and Body nodes in the scene graph. The goal of FBA definition is the animation of both realistic and cartoonist characters. Thus, MPEG-4 has defined a large set of parameters and the user can select subsets of these parameters according to the application. MPEG-4 specifies 84 feature points on the neutral face, which provide spatial reference for FAPs definition. The FAP set contains two high-level parameters, visemes and expressions. In particular, the Facial Definition Parameter (FDP) and the Facial Animation Parameter (FAP) set were designed in the MPEG-4 framework to allow the definition of a facial shape and texture, eliminating the need for specifying the topology of the underlying geometry, through FDPs, and the animation of faces reproducing expressions, emotions and speech pronunciation, through FAPs. Viseme definition has been included in the standard for synchronizing movements of the mouth related to phonemes with facial animation [11].

Although FAPs provide all the necessary elements for MPEG-4 compatible animation, we cannot use them for the analysis of expressions from video scenes, due to the absence of a clear quantitative definition framework. In order to measure FAPs in real image sequences, we have to define a mapping between them and the movement of specific FDP feature points (FPs), which correspond to salient points on the human face.

We have implemented a quantitative modeling of FAPs using features labeled as f_i ($i=1..15$). This feature set employs feature points that lie in the facial area and, in Man Machine Interaction environments, can be automatically detected and tracked. It consists of distances between protuberant points in the facial area. Some of these points are constant during expressions and can be used as reference points; distances between these points are used for normalization purposes [14-16].

2) Facial Feature Extraction

The facial feature extraction scheme used in the system proposed in the framework of the IST ERMIS project is based on an hierarchical, robust scheme, coping with large variations in the appearance of diverse subjects, as well as of the same subject in various instances within real video sequences. Soft *a priori* assumptions are made on the pose of the face or the general location of the features in it. Gradual revelation of information concerning the face is supported under the scope of optimization in each step of the hierarchical scheme, producing *a posteriori* knowledge about

it and leading to a step-by-step visualization of the features in search. This comes in contrast with the basic perspective of other solutions proposed in literature [5], [6], [2], [8], [7], even for emotion recognition [3], [4], which use specific feature representation models or presume an upright position of the face.

Face detection is performed first, as was described in section II above. Following this, primary facial features, such as eyes, mouth and nose, are dealt as major discontinuities on the segmented, arbitrarily rotated face. In the first step of the method, the system performs an optimized segmentation procedure. The initial estimates of the segments, also called seeds, are approximated through min-max analysis and refined through the maximization of a conditional likelihood function. Enhancement is needed so that closed objects will occur and part of the artifacts will be removed. Seed growing is achieved through expansion, utilizing chromatic and value information of the input image. The enhanced seeds form an object set, which reveals the in-plane facial rotation through the use of active contours [13] applied on all objects of the set, which is restricted to a finer set, where the features and MPEG-4 feature points are finally labeled according to an error minimization criterion [12].

In a simplified version of this approach, morphological operations (erosions and dilations), taking into account symmetries, are used to define first the most probable blobs within the facial area to include the eyes and the mouth. Searching through gradient filters over the eyes and between the eyes and mouth provide estimates of the eyebrow and nose positions. Based on the detected facial feature positions, MPEG-4 feature points are then computed and evaluated.

The main problems that facial feature extraction approaches are facing are due to image variations, specifically lighting conditions, low camera precision, orientation and pose and partial occlusions. The method we have developed in ERMIS can cope with large variations in the appearance of diverse subjects, as well as of the same subject in various instances within real video sequences, being also robust to face pose and partial inclusion. However, the major issue of illumination variations is not being effectively tackled. That is why, in this paper, we are proposing a new approach, where the basic method is combined with a (post-processing) neural network subsystem, that evaluates the obtained results, especially in the eye regions which are the most difficult to accurately extract when the above problems exist, and refines them adapting to the specific lighting and capturing conditions.

A multilayer perceptron architecture has been used to evaluate the results provided by the basic method and to adapt its a-priori knowledge about the features of the eye region so as to fit with the properties of the extracted (with high confidence) eye subregions. The features that have been used as inputs to this neural network are thirteen in total and have been acquired from the following pixel-based operation on the image. The images were firstly converted to YCbCr color space and the Y, Cb, and Cr values of the pixel were the first to include in the feature vector. The remaining features were evaluated by applying discrete cosine transformation to the Y (luminance) channel of an 8×8 block of pixels, having as center the specific pixel, and selecting the first ten cosine coefficients of the DCT-transformed resulting block. The multi-layer perceptron consisted of 2 hidden layers, the first of which had 20 neurons, while the output layer consisted of a single neuron. The network was trained using the Levenberg-Marquardt gradient descent method with momentum. The output of the network is shown using a mask on the detected eye region.

3) Experimental Results

Figure 3 below shows a characteristic frame from an input sequence. After face detection, the primary facial features are shown in Figure 4, showing the initially detected blobs, which include eyes and mouth, with Figure 5 showing the feature point estimates, including the eyebrow and nose positions. In Figure 6 the horizontal axis indicates the FAP number, while the vertical axis shows the corresponding FAP values estimated through the selected features.

B. Recognition of basic emotional states

1) Facial Expression Analysis Subsystem

Let us consider a 15-element length feature vector $\underline{f}$, to be the input to the emotion analysis sub-system. The particular values of $\underline{f}$ can be rendered to FAP values resulting in an input vector $\underline{G}$. The elements of $\underline{G}$ express the observed values of the corresponding involved FAPs.

In the following we use expression profiles so as to capture variations of FAPs [9], [10].

Let $X_{i,j}^{(k)}$ be the range of variation of FAP F_j involved in the *k-th* profile $P_i^{(k)}$ of emotion *i*. If $c_{i,j}^{(k)}$ and $s_{i,j}^{(k)}$ are the middle point and length of interval $X_{i,j}^{(k)}$ respectively, then we describe a fuzzy class $A_{i,j}^{(k)}$ for F_j, using the membership function $\mu_{i,j}^{(k)}$ shown in Figure 7. Let also $\Delta_{i,j}^{(k)}$ be the set of classes $A_{i,j}^{(k)}$ that correspond to profile $P_i^{(k)}$; the beliefs $p_i^{(k)}$ and b_i corresponding to profile $P_i^{(k)}$ and emotion *i* respectively, are computed through the following equations:

$$p_i^{(k)} = \prod_{A_{i,j}^{(k)} \in \Delta_{i,j}^{(k)}} r_{i,j}^{(k)} \quad \text{and} \quad b_i = \max_k (p_i^{(k)}), \qquad (1)$$

where $r_{i,j}^{(k)} = \max\{g_i \cap A_{i,j}^{(k)}\}$ expresses the *relevance* of the *i*-th element of the input feature vector with respect to class $A_{i,j}^{(k)}$. Actually $\underline{g} = A'(\underline{G}) = \{g_1, g_2, ...\}$ is the fuzzified input vector resulting from a *singleton* fuzzification procedure.

Figure 3. The original frame from the input sequence

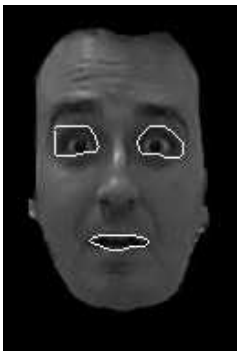


Figure 4 Detected primary facial features

Figure 5 Detected facial features

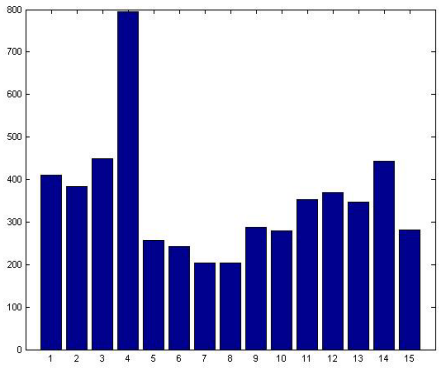


Figure 6 Estimated FAP values

If a hard decision about the observed emotion has to be made then the following equation is used:

$$q = \arg\max_i b_i \,, \qquad (2)$$

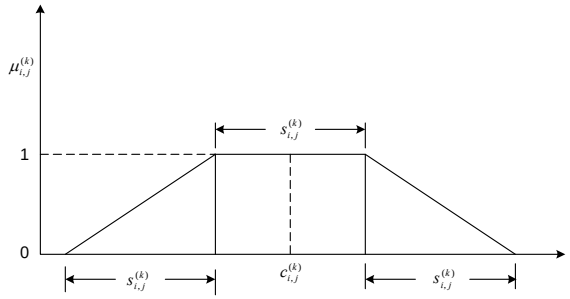


Figure 7. The form of membership functions

The various emotion profiles correspond to the fuzzy intersection of several sets and are implemented through a *τ-norm* of the form $t(a,b)=a \cdot b$. Similarly the belief that an observed feature vector corresponds to a particular emotion results from a fuzzy union of several sets through an *σ-norm* which is implemented as $u(a,b)=\max(a,b)$.

The proposed facial expression analysis system is shown in Figure 8. It provides as result the possible emotions of the user, each accompanied by a degree of belief [16].

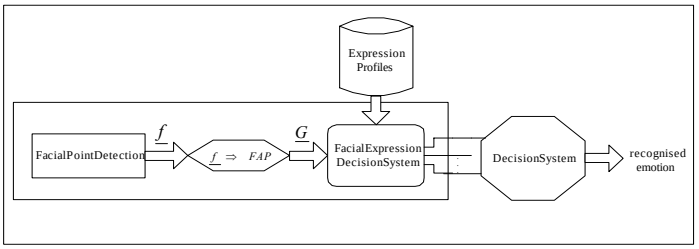
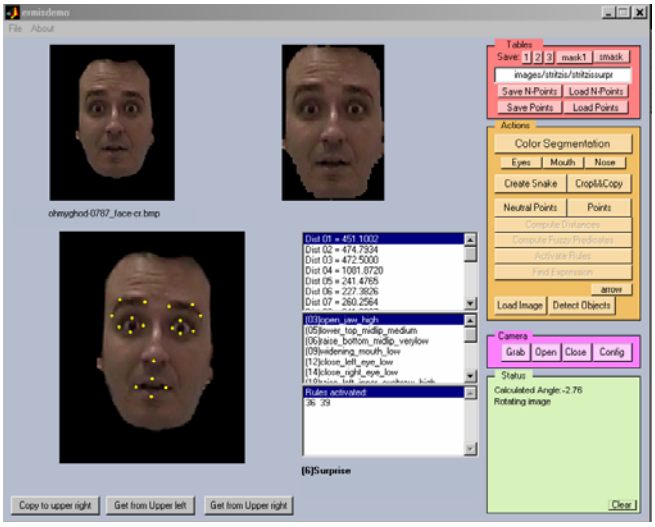


Figure 8. Block diagram of the proposed scheme

2) Experimental Results

In the system interface shown in Figure 9 (a) below, one can observe an example of the calculated FP distances, the profiles selected by the facial expression analysis subsystem and the recognized emotion (“surprise”). In Figure 9(b) classification of the emotional state of the user in one of 4 quadrants (positive/negative: +/-, and active/passive: +/-) is shown together with the detected FPs.

In the following we present results that illustrate the success of the proposed approach for extraction of facial features from the common set of generated data that were of low quality, since they were captured by an analog camera, with illumination problems. The above-described neural network based approach was able to effectively handle this.



(a)

+/+

-/+

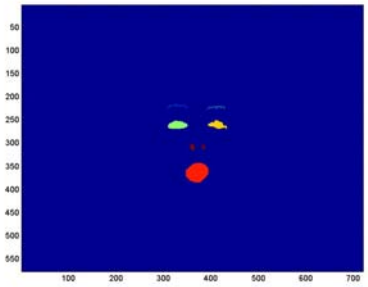
-/-

-/+

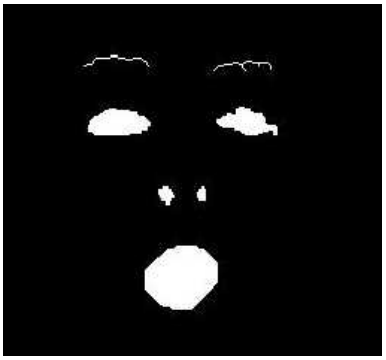
(b)

Figure 9 (a) Facial expression analysis interface (b) Emotional states, FPs and the classifications in the 4 quadrants of the emotional space

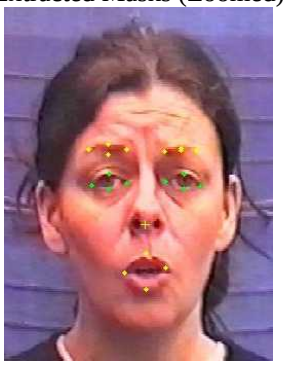
Original Image



Extracted Masks



Extracted Masks (Zoomed)



Feature Points

Figure 10 Extraction of facial features and FPs

Figure 10 shows a frame of the sequences, the obtained features/masks, as well as the feature points that were extracted by the proposed approach. The groundtruth data was generated by experts who have selected (drawn) the true mask locations on more than 100 frames.

Figure 11 shows the superimposed true and auto-detected masks. From the multiplication of the true and auto-detected masks we get the true "common" mask, which represents the true-feature inclusion percentage of the auto-detected mask (brown). By subtracting this common mask from the auto-detected mask, we obtain the non-feature percentage of the auto-detected mask. Figure 12 shows the detected versus the true facial feature masks (eyes, eyebrows, mouth, nose).

The above results indicate that the performance of the proposed feature and FP extraction points is very good, even in difficult cases such as the ones listed above.

IV. Conclusions – Future Work

In this work we have fully adopted the ISO MPEG-4 standard with respect to the FAPs, FDPs and the computed feature parameters. In this way, the analysis results to be produced will be compatible with MPEG-4 based animation, so that it will be straightforward to include or combine the developed system in HCI applications, where a user can interact with an emotionally responding avatar or virtual character. It is expected that the results can further be improved by fitting a 3D model of a face to the 2D image data.

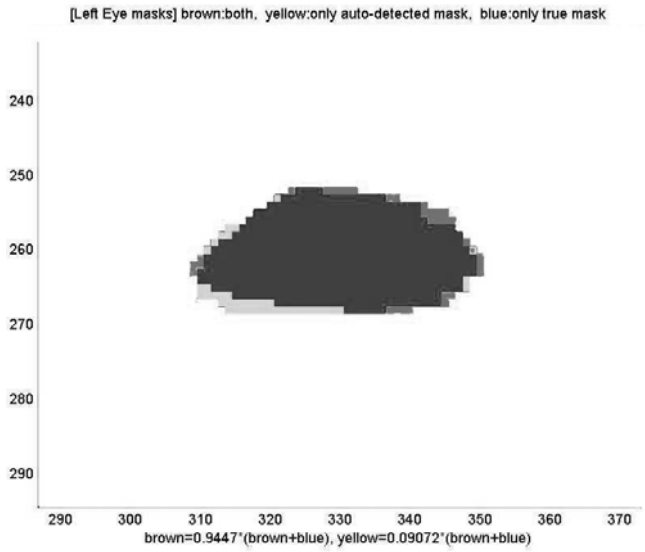


Figure 11. Superimposed true and auto-detected masks are shown.

(a) (b)

Figure 12. (a) Detected feature masks. (b) True (ground truth) feature masks.

Acknowledgement

This work has been supported by the EU funded ERMIS project (IST-2000-29319).

References

[1] R. Fransens, J. De Prins, and L. Van Gool, "SVM-based Nonparametric Discriminant Analysis, An Application to Face Detection", *In Proc. of Intern. Conference on Computer Vision, 2003.*

[2] J. Ahlberg, "An Optimisation Approach to facial feature extraction", *Proc. of IEEE Workshop on Real-Time Analysis and Tracking of Face and Gesture in Real-Time Systems*, Kerkyra, Greece, September 1999.

[3] R. Chellappa, C. L. Wilson and S. Sirohey, "Human and Machine Recognition of Faces: A Survey", *Proc .of the IEEE*, vol. 83, no. 5, pp. 705-740, May 1995.

[4] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, J. G. Taylor, "Emotion Recognition in Human Computer Interaction", *IEEE Signal Processing Magazine*, vol.18, Jan.2001.

[5] P. Eisert and B. Girod, "Model-Based Estimation of Facial Expression Parameters from Image Sequences", *Proc. of IEEE International Conference on Image Processing*, 1997.

[6] Y. Tian, T. Kanade and J. F. Cohn, "Multi-State Based Facial Feature Tracking and Detection", tech. report CMU-RI-TR-99-18, Robotics Institute, Carnegie Mellon University, Aug.1999.

[7] K. Sobottka and I. Pitas, "Looking for Faces and Facial Features in Color Images", Pattern Recognition and Image Analysis: Advances in Mathematical Theory and Applications, vol. 7, no. 1, 1997.

[8] D. Metaxas, "Deformable Model and HMM-Based Tracking, Analysis and Recognition of Gestures and Faces", *Proc. of IEEE Workshop on Real-Time Analysis and Tracking of Face and Gesture in Real-Time Systems*, Kerkyra, Greece, September 1999.

[9] D.K. Panjwani, G. Healey, "Markov Random Field Models for Unsupervised Segmentation of Textured Color Images", *IEEE Trans.* on *PAMI*, vol. 17, no. 10, pp. 939-954, October 1995.

[10] A. Papoulis, "Probability, Random Variables, and Stochastic Processes", McGraw-Hill, Singapore, 3rd Edition, pp. 86-123, 1991.

[11] M. Preda & F. Prêteux, "Advanced animation framework for virtual characters within the MPEG-4 standard", *Proc. of the Intl Conference on Image Processing*, Rochester, NY, 2002.

[12] Y. Votsis, N. Drosopoulos and S. Kollias, "Facial feature segmentation: a modularly optimal approach on real sequences" *Signal Processing: Image Communication,* vol. 18 , no 1, pp. 67-89, 2003.

[13] G. Tsechpenakis, N. Tsapatsoulis and S. Kollias, "Probabilistic Boundary-Based Contour Tracking with Snakes in Natural Cluttered Video Sequences", Intern. Journal Image and Graphics: Special Issue 'Deformable Models: Image Analysis-Pattern Recognition', to appear.

[14] A. Raouzaiou, N. Tsapatsoulis, K. Karpouzis and S. Kollias, "Parameterized facial expression synthesis based on MPEG-4," *Eurasip Journal on Applied Signal Processing,* Vol. 2002, No 10, pp. 1021-1038, 2002.

[15] N. Tsapatsoulis, A. Raouzaiou, S. Kollias, R. Cowie, E. Douglas-Cowie, "Emotion Recognition & Synthesis based on MPEG-4 FAPs", in *MPEG-4 Facial Animation*, John Wiley & Sons, UK, 2002.

[16] K. Karpouzis, A. Raouzaiou, A. Drosopoulos, S. Ioannou, T. Balomenos, N. Tsapatsoulis and S. Kollias, "Facial Expression and Gesture Analysis for Emotionally-rich Man-machine Interaction", N. Sarris, M. Strintzis, (eds.), "3D Modeling and Animation: Synthesis and Analysis Techniques", *Idea Group Publ*., to appear.