ADAPTATION OF EXPRESSION ANALYSIS BASED ON EVALUATION PRINCIPLES

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

ABSTRACT: This paper examines the problem of facial expression analysis in an evaluation-adaptation framework. In this paper, we are proposing a new approach, where the basic method is combined with a (post-processing) evaluation subsystem, which evaluates the obtained features and defines facial features that are to be refined. Anatomical constraints related to both 2-D and 3-D facial models are used for this purpose.
Experimental results are presented, obtained in the framework of the ERMIS IST project, which illustrate the success of the proposed approach for extraction of facial features and expression analysis from low quality real life data.

KEYWORDS: face detection, feature extraction, facial expressions, MPEG-4 facial definition parameters, activation, feature points.

INTRODUCTION

This paper examines the problem of facial expression analysis in an evaluation-adaptation framework. In particular, facial expression analysis is tackled, as in former publications of ours [1], [2], [3], [4], through extraction of MPEG-4 feature points and computation of corresponding facial animation parameters. However, problems may occur due to image variations, specifically lighting conditions, low camera precision, orientation and pose and partial occlusions. In such cases, feature extraction provides partially inaccurate results, which may lead the expression analysis module in erroneous conclusions.
In this paper, we are proposing a new approach, where the basic method is combined with a (post-processing) evaluation subsystem, which evaluates the obtained features and defines facial features that are to be refined. Anatomical constraints related to both 2-D and 3-D facial models are used for this purpose. The extracted conclusions are used to create the necessary input-output data for retraining a neural network that incorporates the a-priori knowledge for facial feature extraction.
Experimental results are presented, obtained in the framework of the ERMIS IST project, which illustrate the success of the proposed approach for extraction of facial features and expression analysis from low quality real life data generated with illumination problems.

FACE DETECTION AND FACIAL FEATURE EXTRACTION

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. These restrictions are being eventually revoked in the literature, since authors deal more and more with realistic environments, while keeping in mind pioneering works in the field.
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 [7] 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.

The facial feature extraction scheme used in the system proposed in the framework of the IST ERMIS project is based on a 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 [9], [10], even for emotion recognition [8] which use specific feature representation models or presume an upright position of the face.

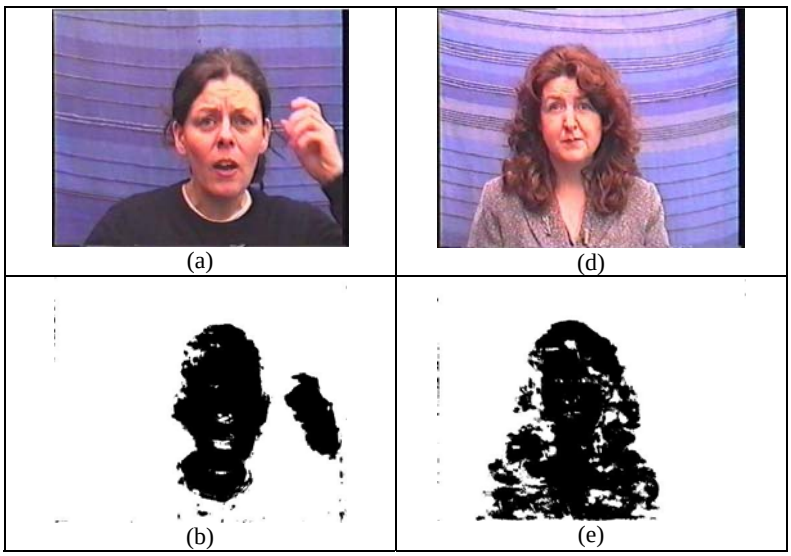
Following the face detection, primary facial features, such as eyes, mouth and nose, are dealt as major discontinuities on the segmented, arbitrarily rotated face. In the first step of the method, the system performs an optimized segmentation procedure. The initial estimates of the segments, also called 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 [11] 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 [1].

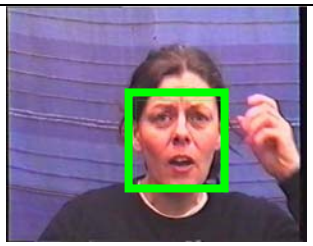
In a simplified version of this approach, morphological operations (erosions and dilations), taking into account symmetries, are used to define first the most probable blobs within the facial area to include the eyes and the mouth. Searching through gradient filters over the eyes and between the eyes 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. To further cope with illumination variations, the 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.

EXPERIMENTAL RESULTS

Figure 1 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.





(c)

(f)

Figure 1: Face detection example ((a), (d): original image, (b), (e): skin color detection, (c), (f): detected face)

Figure 2 shows a characteristic frame from the "hands over the head" sequence. After face detection, the primary facial features are shown in Figure 3. Figure 4 shows the initially detected blobs, which include face and mouth and Figure 5 shows the estimates of the eyebrow and nose positions.

Figure 2: The original frame from the input sequence

Figure 3: Detected primary facial features

Figure 4: The apex of the expression

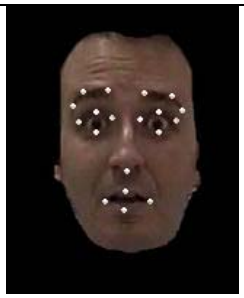


Figure 5: Detected facial features

ANATOMICAL CONSTRAINTS-ANTHROPOMETRY

Anatomical constraints related to both 2-D and 3-D facial models are used in order to evaluate the results obtained from the feature extraction. We compare the distances of Figure 6, [12] with the distances calculated on our test data. The measured distances are normalized by division with Distance 7 (Fig.6), i.e. the distance between the inner corners of left and right eye, both points the human cannot move.

Comparing the measured distances with the range of official measurements, our system can decide if the detected feature points can be accepted. If our measurement is by far out of the normal range of measured distances, the system will ignore the current feature extraction as not accurate.

Figure 6 shows the equivalent distances, officially measured by US Army.

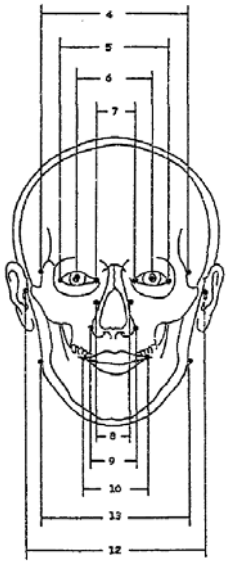


Figure 6: Equivalent measured distances

EXPERIMENTAL RESULTS

Distances DA5n, DA10n correspond to the distances in figures normalized by division with distance DA7:

(DA5n=DA5/DA7,DA10n=DA10/DA7). (1)

DAewn is the eye width calculated from DA5 and DA7:

DAewn=((DA5-DA7)/2)/DA7 (2).

We have measured the distances in Figures 7 and 8 and the measurements are presented in Table I. In Figure one can observe that there is an error in the left eye detection, fact reflected to Dew_rn.

Figure 7

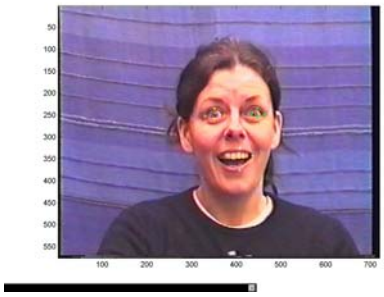


Figure 8

	D5n	DA5n_min	DA5n_max	D10n	DA10n_min	DA10n_max	Dew_ln	Dew_rn	DAewn_min	DAewn_max
Figure 7	2.723	2.517	3.349	1.277	1.031	1.515	0.830	0.894	0.840	1.077
Figure 8	2.129	2.517	3.349	0.919	1.031	1.515	0.677	0.452	0.840	1.077

Table I: Measured distances of the figures

EVALUATION

EVIDENCE THEORY

Evidence Theory [5] is a model for reasoning under uncertainty, alternative to Bayesian reasoning. Similarly to probability theory, a set Ω is considered, which is called a frame of discernment or universal set. Ω contains all mutually contradicting hypotheses. Unlike probability theory, the mass of probability is not distributed among elements of Ω, but rather among subsets of it.

DECISION ANALYSIS IN PATTERN RECOGNITION BASED ON EVIDENCE THEORY

In [6], Denouex presents a method for pattern classification based on evidence theory. An object is given, and the classifier must decide to which classes it belongs, given the evidence of its features. The frame of discernment Ω is the set of possible classes, to which the object can be assigned. The method, which belongs to the k-nearest neighbors methods, produces a basic belief assignment with focal elements the singletons $\{\omega_q\} \subseteq \Omega, \quad q = 1,\ldots,M$ and the universal set Ω. Moreover, the decision procedure is analyzed, which is based on pignistic probability. The object is assigned to the class with the highest pignistic probability, unless the two highest belief assignments are almost equal (the object is rejected due to impossibility of distinguish) or too great a part of the belief has been assigned to the universal set, which means that the object is too far from the known classes (assignment to an unknown class).

EVALUATION OF HUMAN FACE DETECTION

One can create a belief assignment for each component of the face, and perform an evaluation stage to decide whether to verify detected faces or to reject them. Detection of the face, which is a composite object, depends not only to detection of its parts, but also of a correct relation between them which, in this case is a geometrical one. We have produced an initial implementation of the theory for evaluating faces detected in the first step of our approach, by examining:

- Evidence about the class of the eyes (first object, composed of left/right eye regions) and mouth (second object)
- Evidence about the relation among them (mouth region located below eye's one)

We computed the pignistic probability distribution, which is a joint probability distribution and through it, the initial results provided positive evaluations of facial detection in cases when facial features satisfied their geometrical position constraints and negative in case of viblation of them.

EXPRESSION ANALYSIS

FAP AND FDP LOCALIZATION

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 [13].
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 [2], [3], [4].

RECOGNITION OF BASIC EMOTIONAL STATES

Let us consider a 15-element length feature vector $\bar{f}$, to be the input to the emotion analysis sub-system. The particular values of $\bar{f}$ can be rendered to FAP values resulting in an input vector $\bar{G}$. The elements of $\bar{G}$ express the observed values of the corresponding involved FAPs.
In the following we use expression profiles so as to capture variations of FAPs [14]. 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 $u(a,b)=\max(a,b)$.

CONCLUSIONS

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. We have

introduced an evaluation approach based on anthropometric models and measurements as well as on evidence theory, which can be used in order to provide adaptation of expression analysis in real life environments. First results are very promising, Further work is being carried out in the framework of IST project ERMIS.

REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] G. Shafer, A mathematical theory of evidence, Princeton University Press, 1976

[6] T. Denoeux, A k-nearest neighbor classification rule based on Dempster – Shafer theory, 1995, http://citeseer.nj.nec.com/denoeux95knearest.html

[7] 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.*

[8] 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, Jan.2001.

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

[10] 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.

[11] G. Tsechpenakis, N. Tsapatsoulis and S. Kollias, "Probabilistic Boundary-Based Contour Tracking with Snakes in Natural Cluttered Video Sequences", International Journal of Image and Graphics: Special Issue on Deformable Models for Image Analysis and Pattern Recognition, to appear, 2003.

[12] J. Young, "Head and Face Anthropometry of Adult U.S. Civilians", Final Report, July 1993.

[13] 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.

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