FACIAL EXPRESSION ANALYSIS

A. Raouzaiou[1], S. Ioannou[1], K. Karpouzis[1], S. Kollias[1], R. Cowie[2]
[1] School of Electrical and Computer Engineering
National Technical University of Athens,
Heroon Polytechniou 9, 157 80 Zographou, Greece
Phone: +30-210-7722491, Fax: +30-210-7722492
email: {araouz, sivann, kkarpou}@image.ntua.gr, stefanos@cs.ntua.gr
[2] Department of Psychology
Queen’s University of Belfast
Northern Ireland, United Kingdom
email: r.cowie@qub.ac.uk

ABSTRACT: There has recently been high interest in affective computing, especially in interfaces which can analyse their users’ emotional state. Automatic emotion recognition in faces is a hard problem, requiring a number of pre-processing steps which attempt to detect or track the face, to locate characteristic facial regions such as eyes, mouth and nose on it, to extract and follow the movement of facial features, e.g., characteristic points in these regions, or model facial gestures using anatomic information about the face.
In this work we present a methodology for analysing both primary and intermediate expressions. This is performed through a system which, after a skin color segmentation and the extraction of the face and of the facial points, translates FP movements into FAPs and reasons on the latter to recognize the underlying emotion in facial video sequences.
The developments described in the current work are being extended and validated in the framework of the IST ERMIS project.

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

INTRODUCTION

Current information processing and visualization systems are capable of offering advanced and intuitive means of receiving input and communicating output to their users. Man-Machine Interaction (MMI) systems give the opportunity to less technology-aware individuals, as well as handicapped people, to use computers more efficiently and thus overcome related fears and preconceptions. Besides this, most emotion-related facial gestures are considered to be universal, in the sense that they are recognized along different cultures. Therefore, the introduction of an “emotional dictionary” that includes descriptions and perceived meanings of facial expressions, so as to help infer the likely emotional state of a specific user, can enhance the affective nature [7] of MMI applications.
In this paper, we present a systematic approach to analyzing emotional cues from user facial expressions. In the first section, we provide an overview of affective analysis of facial expressions, supported by psychological studies describing emotions as discrete points or areas of an “emotional space”. The following sections describe the facial feature extraction which is used in our system and provide algorithms and experimental results from the analysis of facial expressions in video sequences. The motion of tracked feature points is translated to MPEG-4 FAPs, which describe their observed motion in a higher-level manner.

REPRESENTATION OF EMOTION

The obvious goal for emotion analysis applications is to assign category labels that identify emotional states. However, labels as such are very poor descriptions, especially since humans use a daunting number of labels to describe emotion. Therefore we need to incorporate a more transparent, as well as continuous representation, that matches closely our conception of what emotions are or, at least, how they are expressed and perceived.

Activation-emotion space [8], [5] is a representation that is both simple and capable of capturing a wide range of significant issues in emotion. It rests on a simplified treatment of two key themes:

- *Valence*: The clearest common element of emotional states is that the person is materially influenced by feelings that are “valenced”, i.e. they are centrally concerned with positive or negative evaluations of people or things or events; the link between emotion and valencing is widely agreed.
- *Activation leve*l: Research has recognized that emotional states involve dispositions to act in certain ways. A basic way of reflecting that theme turns out to be surprisingly useful. States are simply rated in terms of the associated activation level, i.e. the strength of the person’s disposition to take some action rather than none.

A surprising amount of emotional discourse can be captured in terms of activation-emotion space. Perceived full-blown emotions are not evenly distributed in activation-emotion space; instead they tend to form a roughly circular pattern. From that and related evidence, work presented in [6] shows that there is a circular structure inherent in emotionality. In this framework, identifying the center as a natural origin has several implications. Emotional strength can be measured as the distance from the origin to a given point in activation-evaluation space. The concept of a full-blown emotion can then be translated roughly as a state where emotional strength has passed a certain limit. An interesting implication is that strong emotions are more sharply distinct from each other than weaker emotions with the same emotional orientation. A related extension is to think of primary or basic emotions as cardinal points on the periphery of an emotion circle. Plutchik [6] has offered a useful formulation of that idea, the ‘emotion wheel’.

FACIAL EXPRESSION ANALYSIS

FACIAL FEATURES RELEVANT TO EXPRESSION ANALYSIS

Facial analysis includes a number of processing steps which attempt to detect or track the face, to locate characteristic facial regions such as eyes, mouth and nose on it, to extract and follow the movement of facial features, such as characteristic points in these regions, or model facial gestures using anatomic information about the face.

Most of the above models are based on a well-known system for describing “all visually distinguishable facial movements” called the Facial Action Coding System (FACS) [4]. The FACS model has inspired the derivation of facial animation and definition parameters in the framework of the ISO MPEG-4 standard [3]. 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. By monitoring facial gestures corresponding to FDP and/or FAP movements over time, it is possible to derive cues about user’s expressions and emotions. Various results have been presented regarding classification of archetypal expressions of faces, mainly based on features or points mainly extracted from the mouth and eyes areas of the faces.

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.

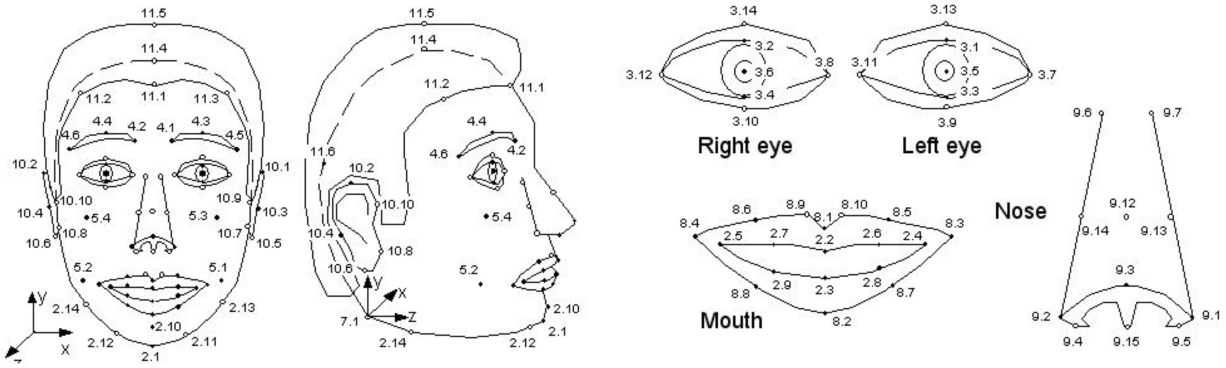


Figure 1: FDP feature points (adapted from [3])

Table I provides the quantitative modeling of FAPs that we have implemented using the features labeled as f_i ($i=1..15$) [2]. 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, noted as *s(x,y)*, between protuberant points, *x* and *y*, corresponding to the Feature Points shown in Figure 1. Some of these points are constant during expressions and can be used as reference points; distances between these points are used for normalization purposes [9].

FAP name	**Feature for the description**	**Utilized feature**
Squeeze_l_eyebrow (F_{37})	$D_{1}=s(4.5,3.11)$	$f_{1=}\ D_{1\text{-NEUTRAL}} - D_{1}$
Squeeze_r_eyebrow (F_{38})	$D_{2}=s(4.6,3.8)$	$f_{2=}\ D_{2\text{-NEUTRAL}} - D_{2}$
Lower_t_midlip (F_{4})	$D_{3}=s(9.3,8.1)$	$f_{3=}\ D_{3} - D_{3\text{-NEUTRAL}}$
Raise_b_midlip (F_{5})	$D_{4}=s(9.3,8.2)$	$f_{4=}\ D_{4\text{-NEUTRAL}} - D_{4}$
Raise_l_i_eyebrow (F_{31})	$D_{5}=s(4.1,3.11)$	$f_{5=}\ D_{5} - D_{5\text{-NEUTRAL}}$
Raise_r_i_eyebrow (F_{32})	$D_{6}=s(4.2,3.8)$	$f_{6=}\ D_{6} - D_{6\text{-NEUTRAL}}$
Raise_l_o_eyebrow (F_{35})	$D_{7}=s(4.5,3.7)$	$f_{7=}\ D_{7} - D_{7\text{-NEUTRAL}}$
Raise_r_o_eyebrow (F_{36})	$D_{8}=s(4.6,3.12)$	$f_{8=}\ D_{8} - D_{8\text{-NEUTRAL}}$
Raise_l_m_eyebrow (F_{33})	$D_{9}=s(4.3,3.7)$	$f_{9=}\ D_{9} - D_{9\text{-NEUTRAL}}$
Raise_r_m_eyebrow (F_{34})	$D_{10}=s(4.4,3.12)$	$f_{10=}\ D_{10} - D_{10\text{-NEUTRAL}}$
Open_jaw (F_{3})	$D_{11}=s(8.1,8.2)$	$f_{11=}\ D_{11} - D_{11\text{-NEUTRAL}}$
close_t_l_eyelid (F_{19}) – *close_b_l_eyelid* (F_{21})	$D_{12}=s(3.1,3.3)$	$f_{12=}\ D_{12} - D_{12\text{-NEUTRAL}}$
close_t_r_eyelid (F_{20}) – *close_b_r_eyelid* (F_{22})	$D_{13}=s(3.2,3.4)$	$f_{13=}\ D_{13} - D_{13\text{-NEUTRAL}}$
stretch_l_cornerlip (F_{6}) *(stretch_l_cornerlip_o)*(F_{53}) – *stretch_r_cornerlip* (F_{7}) *(stretch_r_cornerlip_o)*(F_{54})	$D_{14}=s(8.4,8.3)$	$f_{14=}\ D_{14} - D_{14\text{-NEUTRAL}}$
squeeze_l_eyebrow (F_{37}) AND *squeeze_r_eyebrow* (F_{38})	$D_{15}=s(4.6,4.5)$	$f_{15=}\ D_{15\text{-NEUTRAL}} - D_{15}$

Table I: Quantitative FAP modeling: (1) **s(x,y)** is the Euclidean distance between the FPs, (2) $\mathbf{D_{i\text{-NEUTRAL}}}$ refers to the distance $\mathbf{D_{i}}$ when the face is its in neutral position

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.

The facial feature extraction scheme used in the system proposed in this chapter 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, we have recently developed [10]. 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.

Face detection is performed first through detection of skin segments or blobs, merging of them based on the probability of their belonging to a facial area, and identification of the most salient skin color blob or segment. 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 applied on all objects of the set, which is restricted to a finer set, where the features and feature points are finally labeled according to an error minimization criterion.

An efficient implementation of the scheme has been developed in the framework of the IST ERMIS project (www.image.ntua.gr/ermis). Following face detection, 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, feature points are computed and evaluated.

FACIAL EXPRESSION ANALYSIS SYSTEM

The facial expression analysis subsystem is the main part of the presented system; gestures are utilized to support the outcome of this subsystem.

Let us consider as input to the emotion analysis sub-system a 15-element length feature vector $\underline{f}$ that corresponds to the 15 features f_i shown in Table I. The particular values of $\underline{f}$ can be rendered to FAP values as shown in the same table resulting in an input vector $\underline{G}$. The elements of $\underline{G}$ express the observed values of the corresponding involved FAPs.

Expression profiles are also used to capture variations of FAPs [9]. For example, the range of variations of FAPs for the expression “surprise” are shown in Table II.

Surprise ($P_{Su}^{(0)}$)	$F_3 \in [569,1201]$, $F_5 \in [340,746]$, $F_6 \in [-121,-43]$, $F_7 \in [-121,-43]$, $F_{19} \in [170,337]$, $F_{20} \in [171,333]$, $F_{21} \in [170,337]$, $F_{22} \in [171,333]$, $F_{31} \in [121,327]$, $F_{32} \in [114,308]$, $F_{33} \in [80,208]$, $F_{34} \in [80,204]$, $F_{35} \in [23,85]$, $F_{36} \in [23,85]$, $F_{53} \in [-121,-43]$, $F_{54} \in [-121,-43]$
$P_{Su}^{(1)}$	$F_3 \in [1150,1252]$, $F_5 \in [-792,-700]$, $F_6 \in [-141,-101]$, $F_7 \in [-141,-101]$, $F_{10} \in [-530,-470]$, $F_{11} \in [-530,-470]$, $F_{19} \in [-350,-324]$, $F_{20} \in [-346,-320]$, $F_{21} \in [-350,-324]$, $F_{22} \in [-346,-320]$, $F_{31} \in [314,340]$, $F_{32} \in [295,321]$, $F_{33} \in [195,221]$, $F_{34} \in [191,217]$, $F_{35} \in [72,98]$, $F_{36} \in [73,99]$, $F_{53} \in [-141,-101]$, $F_{54} \in [-141,-101]$
$P_{Su}^{(2)}$	$F_3 \in [834,936]$, $F_5 \in [-589,-497]$, $F_6 \in [-102,-62]$, $F_7 \in [-102,-62]$, $F_{10} \in [-380,-320]$, $F_{11} \in [-380,-320]$, $F_{19} \in [-267,-241]$, $F_{20} \in [-265,-239]$, $F_{21} \in [-267,-241]$, $F_{22} \in [-265,-239]$, $F_{31} \in [211,237]$, $F_{32} \in [198,224]$, $F_{33} \in [131,157]$, $F_{34} \in [129,155]$, $F_{35} \in [41,67]$, $F_{36} \in [42,68]$
$P_{Su}^{(3)}$	$F_3 \in [523,615]$, $F_5 \in [-386,-294]$, $F_6 \in [-63,-23]$, $F_7 \in [-63,-23]$, $F_{10} \in [-230,-170]$, $F_{11} \in [-230,-170]$, $F_{19} \in [-158,-184]$, $F_{20} \in [-158,-184]$, $F_{21} \in [-158,-184]$, $F_{22} \in [-158,-184]$, $F_{31} \in [108,134]$, $F_{32} \in [101,127]$, $F_{33} \in [67,93]$, $F_{34} \in [67,93]$, $F_{35} \in [10,36]$, $F_{36} \in [11,37]$

Table II: Profiles for the archetypal emotion surprise

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 $m_{i,j}^{(k)}$ shown in Figure 2. 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 that an observed, through the vector $\underline{G}$, facial state corresponds to profile $P_i^{(k)}$ and emotion *i* respectively, are computed through the following equations:

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

where $r_{i,j}^{(k)} = \max\{g_i \cap A_{i,j}^{(k)}\}$ expresses the *relevance* $r_{i,j}^{(k)}$ 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 [1].

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

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

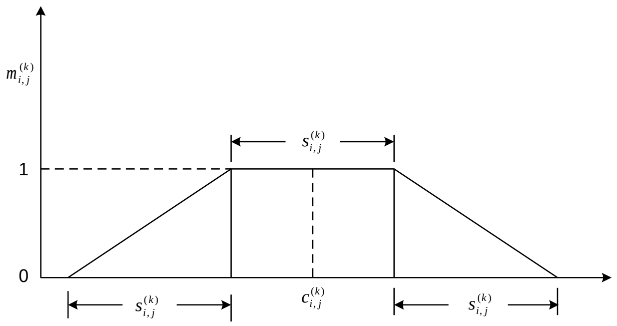
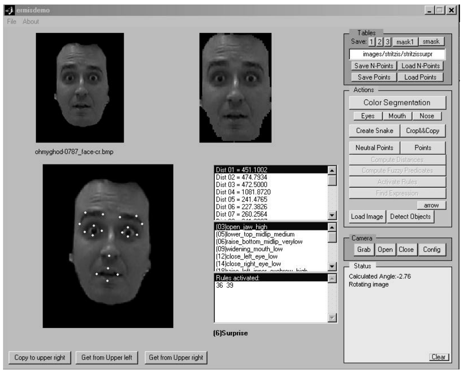


Figure 2: The form of membership functions



images/stritzis/stritzissurpr

ohmyghod-0787_face-cr.bmp

```
Dist 01 = 451.1002
Dist 02 = 474.7934
Dist 03 = 472.5000
Dist 04 = 1081.8720
Dist 05 = 241.4765
Dist 06 = 227.3826
Dist 07 = 260.2564
```

```
(03)open_jaw_high
(05)lower_top_midlip_medium
(06)raise_bottom_midlip_verylow
(09)widening_mouth_low
(12)close_left_eye_low
(14)close_right_eye_low
```

```
Rules activated:
36 39
```

(6)Surprise

Calculated Angle:-2.76
Rotating image

Figure 3: Facial expression analysis interface

The various emotion profiles correspond to the fuzzy intersection of several sets and are implemented through a *t-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 *s-norm* which is implemented as $u(a,b)=\max(a,b)$.

An emotion analysis system has been created as part of the IST ERMIS project (www.image.ntua.gr/ermis). In the system interface shown in Figure 3, one can observe an example of the calculated FP distances, the profiles selected by the facial expression analysis subsystem and the recognized emotion ("surprise"). The proposed facial expression analysis system is shown in Figure 4. The system provides as result the possible emotions of the user, each accompanied by a degree of belief.

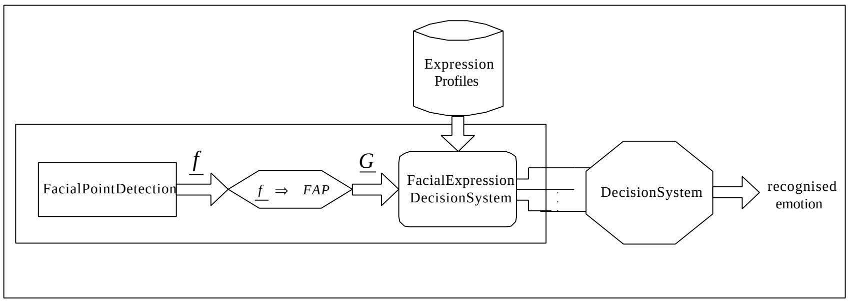


Figure 4: Block diagram of the proposed scheme

EXPERIMENTAL RESULTS

Figure 5 shows a characteristic frame from the "hands over the head" sequence. After skin detection and segmentation, the primary facial features are shown in Figure 6. Figure 7 shows the detected blobs, i.e. face and mouth, and Figure 8 shows the estimates of the eyebrow and nose positions. Figure 9 shows the initial neutral image used to calculate the FP distances. In Figure 10 the horizontal axis indicates the FAP number, while the vertical axis shows the corresponding FAP values estimated through the features stated in the second column of Table I.

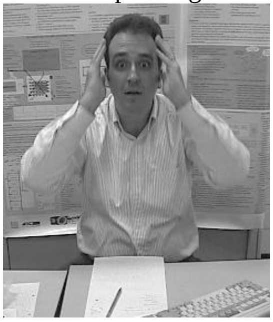


Figure 5: The original frame from the input sequence

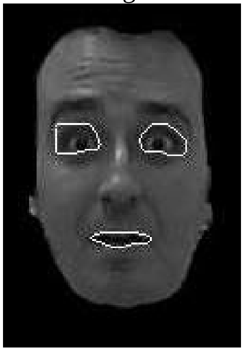


Figure 6: Detected primary facial features

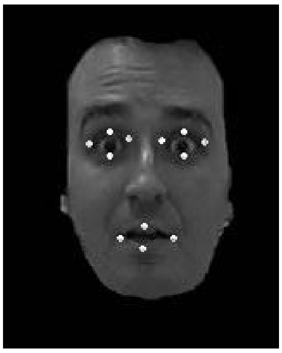


Figure 7: The apex of an expression

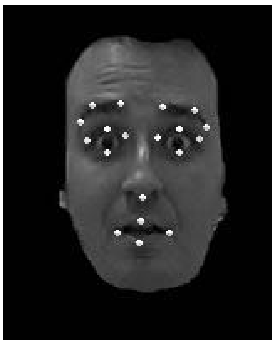


Figure 8: Detected facial features

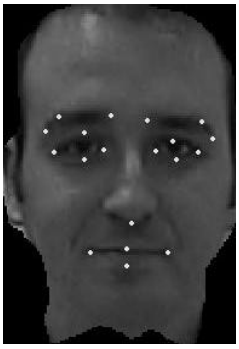


Figure 9: A neutral expression

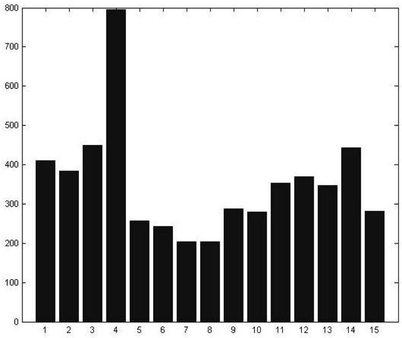


Figure 10: Estimated FAP values for Figure 8

CONCLUSIONS

In this paper we described a holistic approach to emotion modeling and analysis and their applications in MMI applications. Beginning from a symbolic representation of human emotions found in this context, based on their expression via facial expressions, we show that it is possible to transform quantitative feature information from video sequences to an estimation of a user's emotional state. This transformation is based on a fuzzy rules architecture that takes into account knowledge of emotion representation and the intrinsic characteristics of human expression. Input to these rules consists of features extracted and tracked from the input data, i.e. facial features. While these features can be used for simple representation purposes, e.g. animation or task-based interfacing, our approach is closer to the target of affective computing. Thus, they are utilized to provide feedback on the users' emotional state, while in front of a computer. Possible applications include human-like agents, that assist everyday chores and react to user emotions or sensitive artificial listeners that introduce conversation topics and react themselves to specific user cues.

REFERENCES

[1] Klir G.; Yuan, B., 1995, "Fuzzy Sets and Fuzzy Logic, Theory and Application", Prentice Hall, New Jersey.

[2] Karpouzis, K.; Tsapatsoulis, N.; Kollias, S., 2000, "Moving to Continuous Facial Expression Space using the MPEG-4 Facial Definition Parameter (FDP) Set," SPIE Electronic Imaging 2000, San Jose, CA, USA.

[3] Tekalp, A.M.; Ostermann, J., 2000, "Face and 2-D Mesh Animation in MPEG-4", Signal Processing: Image Communication, Vol. 15, pp. 387-421.

[4] Ekman, P.; Friesen, W., 1978, "The Facial Action Coding System", Consulting Psychologists Press, CA.

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

[6] Plutchik, R., 1980, "Emotion: A psychoevolutionary synthesis", Harper and Row, NY, USA.

[7] Picard, R.W., 2000, "Affective Computing", MIT Press, Cambridge, MA..

[8] Whissel, C.M., 1989, "The dictionary of affect in language", Emotion: Theory, research and experience: vol. 4, The measurement of emotions, R. Plutchnik and H. Kellerman (Eds), Academic Press, New York.

[9] Raouzaiou, A.; Tsapatsoulis, N.; Karpouzis, K.; Kollias, S., 2002, "Parameterized facial expression synthesis based on MPEG-4", EURASIP Journal on Applied Signal Processing, Vol. 2002, No. 10, pp. 1021-1038, Hindawi Publishing Corporation.

[10] Votsis, G.; Drosopoulos, A.; Kollias, S., 2003, "A modular approach to facial feature segmentation on real sequences", Signal Processing, Image Communication, vol. 18, pp. 67-89.