EMOTION ANALYSIS IN MAN-MACHINE INTERACTION SYSTEMS

T.Balomenos, A.Raouzaiou, S.Ioannou, A.Drosopoulos, K.Karpouzis, and S.Kollias

Image, Video and Multimedia Systems Laboratory
National Technical University of Athens

ABSTRACT

Facial expression and hand gesture analysis plays a fundamental part in emotionally rich man-machine interaction (MMI) systems, since it employs universally accepted non-verbal cues to estimate the users' emotional state. In this paper, we present a systematic approach to extracting expression related features from image sequences and inferring an emotional state via an intelligent rule-based system. MMI systems can benefit from these concepts by adapting their functionality and presentation with respect to user reactions or by employing agent-based interfaces to deal with specific emotional states, such as frustration or anger.

1. INTRODUCTION

Current information processing and visualization systems are capable of offering advanced and intuitive means of receiving input and communicating output to their users. As a result, Man-Machine Interaction (MMI) systems that utilize multimodal information about their users' current emotional state are presently at the forefront of interest of the computer vision and artificial intelligence communities. Such interfaces 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 and body 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 and body gestures, so as to help infer the likely emotional state of a specific user, can enhance the affective nature [13] of MMI applications.

Despite the progress in related research, our intuition of what a human expression or emotion actually represents is still based on trying to mimic the way the human mind works while making an effort to recognize such an emotion. This means that even though image or video input are necessary to this task, this process cannot come to robust results without taking into account features like speech, hand gestures or body pose. These features provide means to convey messages in a much more expressive and definite manner than wording, which can be misleading or ambiguous. While a lot of effort has been invested in examining individually these aspects of human expression, recent research [10] has shown that even this approach can benefit from taking into account multimodal information.

In this paper, we present a systematic approach to analyzing emotional cues from user facial expressions and hand gestures. Emotions are considered as discrete points or areas of an "emotional space" [10]. In section 2, we provide an overview of affective analysis of facial expressions and gestures. Sections 3 and 4 provide algorithms and experimental results from the analysis of facial expressions and hand gestures in video sequences. In most cases a single expression or gesture cannot help the system deduce a positive decision about the users' observed emotion. As a result, a fuzzy architecture is employed that uses the symbolic representation of the tracked features as input; this concept is described in Section 5. Results of the multimodal affective analysis system are provided in this section, while conclusions and future work concepts are included in Section 6.

2. AFFECTIVE ANALYSIS IN MMI

2.1 Affective facial expression analysis

There is a long history of interest in the problem of recognizing emotion from facial expressions [9], and extensive studies on face perception during the last twenty years [7], [5]. The salient issues in emotion recognition from faces are parallel in some respects to the issues associated with voices, but divergent in others.

In the context of faces, the task has almost always been to classify examples of archetypal emotions. That may well reflect the influence of Ekman and his colleagues, who have argued robustly that the facial expression of emotion is inherently categorical. More recently, morphing techniques have been used to probe states that are intermediate between archetypal expressions. They do reveal effects that are consistent with a degree of categorical structure in the domain of facial expression, but they are not particularly large, and there may be alternative ways of explaining them – notably by considering how category terms and facial parameters map onto activation-evaluation space [6].

2.2 Affective gesture analysis

The detection and interpretation of hand gestures has become an important part of human computer interaction (MMI) in recent years [14]. Sometimes, a simple hand action, such as placing a person's hands over his ears, can pass on the message that he has had enough of what he is hearing; this is conveyed more expressively than with any other spoken phrase. Analyzing hand gestures is a comprehensive task involving motion modeling, motion analysis, pattern recognition, machine learning, and even psycholinguistic studies.

The first phase of the recognition task is choosing a model of the gesture. The mathematical model may consider both the spatial and temporal characteristic of the hand and hand gestures [4]. The approach used for modeling plays a pivotal role in the nature and performance of gesture interpretation. Once the model is decided upon, an analysis stage is used to compute the model parameters from the image features that are extracted from single or multiple video input streams. These parameters constitute some description of the hand pose or trajectory and depend on the modeling approach used. Among the important problems involved in the analysis are those of hand localization, hand tracking [11], [12], [1] and selection of suitable image features. The computation of model parameters is followed by gesture recognition. Here, the parameters are classified and interpreted in the light of the accepted model and perhaps the rules imposed by some grammar. The grammar could reflect not only the internal syntax of gestural commands but also the possibility of interaction of gestures with other communication modes like speech, gaze, or facial expressions.

3. FACIAL EXPRESSION ANALYSIS

3.1 Facial Features Extraction

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.

Although FAPs [8] 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 [15].

The facial feature extraction scheme used in the system proposed in this paper 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 [16]. 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. Primary facial features, such as eyes, mouth and nose, are dealt as major discontinuities on the segmented, arbitrarily rotated face. 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

An efficient implementation of the scheme has been developed in the framework of the IST ERMIS project (www.image.ntua.gr/ermis)..

3.2 Experimental Results

Figure 1 shows a characteristic frame from an image sequence. After skin detection and segmentation, the primary facial features are shown in Figure 2. Figure 3 shows the estimates of the eyes, mouth, eyebrows and nose positions. Figure 4 shows the initial neutral image used to calculate the FP distances.

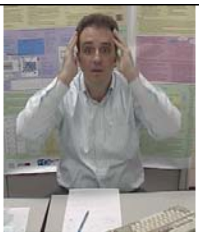

Figure 1: The original frame from the input sequence

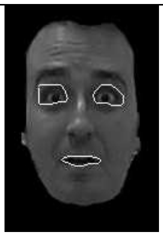

Figure 2: Detected primary facial features

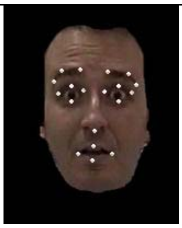

Figure 3: Detected features in the apex of an expression

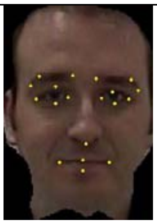

Figure 4: A neutral expression

4. GESTURE ANALYSIS

4.1 Hand detection and tracking

In order to extract emotion-related features through hand movement, we implemented a hand-tracking system. Emphasis was on implementing a near real-time, yet robust enough system for our purposes. The general process involves the creation of *moving skin masks*, namely skin color areas which are tracked between subsequent frames. By tracking the centroid of those skin masks we produce an estimate of the user’s movements. In order to implement a computationally light system, our architecture takes into account a-priori knowledge related to the expected characteristics of the input image. Since the context is MMI applications, we expect to locate the head in the middle area of upper half of the frame and the hand segments near the respective lower corners. In addition to this, we concentrate on the motion of hand segments, given that they are the end effectors of the hand and arm chain and thus the most expressive object in tactile operations.

For each given frame, as in the face detection process, a skin color probability matrix is computed by calculating the joint probability of the Cr/Cb image values (Figure 5). A skin color mask is then obtained from the skin probability matrix with thresholding (Figure 6). Possible moving areas are found by thresholding the difference pixels between the current frame and the next, resulting to the possible-motion mask (Figure 7). This mask does not contain information about the direction or the magnitude of the movement, but is only indicative of the motion and is used to accelerate the algorithm by concentrating tracking only in moving image areas. Both color (Figure 6) and motion (Figure 7) masks contain a large number of small objects due to the presence of noise and objects with color similar to the skin. To overcome this, morphological filtering is employed on both masks to remove small objects. All described morphological operations are carried out with a disk structuring element with a radius of 1% of the image width. The distance transform of the color mask is first calculated and only objects above the desired size are retained. These objects are used as markers for the morphological reconstruction of the initial color mask. The color mask is then closed to provide better centroid calculation. The moving skin mask (msm) is then created by fusing the processed skin and motion masks (sm, mm) through the morphological reconstruction of the color mask using the motion mask as marker. The result of this process, after excluding the head object is shown in (Figure 8). The moving skin mask consists of many large connected areas. For the next frame a new moving skin mask is created, and a one-to-one object correspondence is performed. Object correspondence between two frames is performed on the color mask and is based on object centroid distance for objects of similar (at least 50%) area (Figure 9). In these figures, red markers (crosses) represent the position of the centroid of the detected right hand of the user, while green markers (circles) correspond to the left hand. In the case of hand object merging and splitting, e.g. in the case of clapping, we establish a new matching of the left-most candidate object to the user's right hand and the right-most object to the left hand (Figure 10). Following object matching in the subsequent moving skin masks, the mask flow is computed, i.e. a vector for each frame depicting the motion direction and magnitude of the frame’s objects. The described algorithm is relatively lightweight, allowing a rate of several fps on a usual PC.

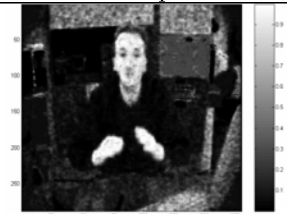

Figure 5: Skin color probability for the input image

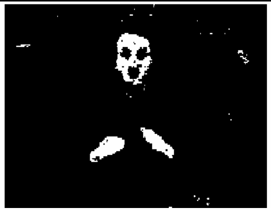

Figure 6: Initial color mask created with skin detection

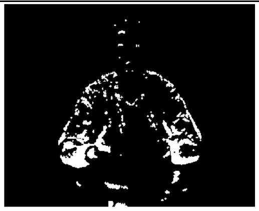

Figure 7: Initial motion mask after pixel difference thresholded to 10% of maximum

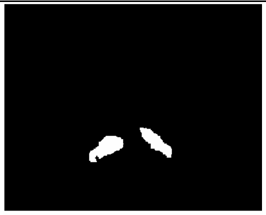

Figure 8: Detected moving hand segments after morphological reconstruction

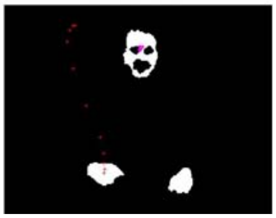

Figure 9: Tracking of one hand object in the “lift of the hand” sequence

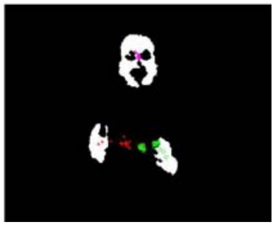

Figure 10: Tracking of both hand objects in the “clapping” sequence

4.2 Gesture Classification using HMMs

The ability of Hidden Markov Models (HMMs) to deal with time sequential data and to provide time scale invariability as well as learning capability makes them an appropriate selection for gesture classification. An excellent study on HMMs can be found in [17]. In Table 1 we present the utilized features that feed (as sequences of vectors) our HMM classifier, as well as the output classes of the HMM classifier.

The recognizer consists of *M* different HMMs corresponding to the modeled gesture classes. In our case, *M=7* as it can be seen in Table 1. We use first order left-to-right models consisting of a varying number (for each one of the HMMs) of internal states $G_{k,j}$ that have been identified through the learning process. For example the third HMM which recognizes low speed on *hand lift* consists of only three states $G_{3,1}$, $G_{3,2}$ and $G_{3,3}$ while more complex gesture classes like the *hand clapping* require as much as eight states to be efficiently modeled by the corresponding HMM.

Features	X_{lh} - X_{rh}, X_f -X_{rh}, X_f -$X_{lh,}$ Y_{lh} - $Y_{rh,}$ Y_f - $Y_{rh,}$ Y_f -Y_{lh} where C_f=(X_f,Y_f) the coordinates of the head centroid, C_{rh}=(X_{rh},Y_{rh}) the coordinates of the right hand centroid, C_{lh}=(X_{lh},Y_{lh}) the coordinates of the left hand centroid
Gesture Classes	hand clapping – high frequency hand clapping – low frequency lift of the hand – low speed lift of the hand – high speed hands over the head – gesture hands over the head – posture italianate gestures

Table 1: a) *Features (inputs to HMM) and* **b)** *Gesture Classes (outputs of HMM)*

4.3 Experimental results

Experiments for testing the recognizing performance of the proposed algorithm were carried out. Gesture sequences of three male subjects, with maximum duration of three seconds, were captured by a typical web-camera at a rate of 10 frames per second. For each one of the gesture classes 15 sequences were acquired, three were used for the initialization of the HMM parameters, seven for training and parameters' re-estimation and five for testing. Each one of the training sequences consisted of approximately 15 frames. The selection of these frames was performed off-line so as to create characteristic examples of the gesture classes. Testing sequences were sub-sampled at a rate of 5 frames per second so as to enable substantial motion to occur. An overall recognition rate of 94,3% was achieved.

From the results obtained we observed a mutual misclassification between "Italianate Gestures" and "Hand Clapping – High Frequency"; this is mainly due to the variations on "Italianate Gestures" across different individuals. Thus, training the HMM classifier on a personalized basis is anticipated to improve the discrimination between these two classes.

5. MULTIMODAL AFFECTIVE ANALYSIS

5.1 Facial Expression Analysis Subsystem

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 subsystem a 15-element length feature vector $\underline{f}$ that corresponds to the 15 features f_i [15]. 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.

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 11. 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: $p_i^{(k)} = \prod_{A_{i,j}^{(k)} \in \Delta_{i,j}^{(k)}} r_{i,j}^{(k)}$ and $b_i = \max_k (p_i^{(k)})$, (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 [3].

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

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

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*·*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 $u(a,b)=\max(a,b)$.

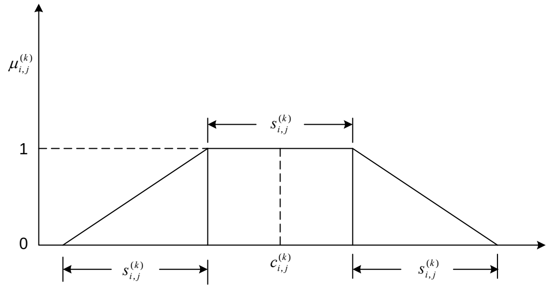

Figure 11: The form of membership functions

An emotion analysis system has been created as part of the IST ERMIS project (www.image.ntua.gr/ermis).

5.2 Affective Gesture Analysis Subsystem

Gestures are utilized to support the outcome of the facial expression analysis subsystem, since in most cases they are too ambiguous to indicate a particular emotion. However, in a given context of interaction, some gestures are obviously associated with a particular expression –e.g. *hand clapping* of high frequency expresses *joy*, *satisfaction*- while others can provide indications for the kind of the emotion expressed by the user. In particular, quantitative features derived from hand tracking, like speed and amplitude of motion, fortify the position of an observed emotion; for example, *satisfaction* turns to *joy* or even to *exhilaration*, as the speed and amplitude of clapping increases. As was mentioned in Section 4, the position of the centroids of the head and the hands over time forms the feature vector sequence that feeds an HMM classifier whose outputs corresponds to particular gesture class. In the following paragraph we describe how the recognized gesture class can be used to provide indications about the occurrence of an emotional state.

Table 2 below shows the correlation between some detectable gestures with the six archetypal expressions.

Emotion	Gesture Class
Joy	*hand clapping-high frequency*
Sadness	*hands over the head-posture*
Anger	*lift of the hand- high speed*
	italianate gestures
Fear	*hands over the head-gesture*
	italianate gestures
Disgust	*lift of the hand- low speed*
	hand clapping-low frequency
Surprise	*hands over the head-gesture*

Table 2: *Correlation between gestures and emotional states*

Given a particular context of interaction, gesture classes corresponding to the same emotional are combined in a "logical OR" form. Table 2 shows that a particular gesture may correspond to more than one gesture classes carrying different affective meaning. For example, if the examined gesture is *clapping*, detection of high frequency indicates *joy*, but a *clapping* of low frequency may express irony and can reinforce a possible detection of the facial expression *disgust*.

In practice, the gesture class probabilities derived by the HMM classifier are transformed to emotional state indicators by using the information of Table 2. Let EI_k be the emotional indicator of emotional state k (k $\in\{1,2,3,4,5,6\}$ corresponds to one of the emotional states presented in Table 2 in the order of appearance, i.e., 1->Joy, 6->Surprise), GCS= {gc$_1$, gc$_2$, …, gc$_N$} be the set of gesture classes recognized by the HMM Classifier (N=7), $GCS^k \subseteq GCS$ be the set of gesture classes related with the emotional state k, and $p(gc_i)$ be the probability of gesture class gc_i obtained from the HMM Classifier. The $EI(k)$ is computed using the following equation: $EI_k = \max_{gc_i \in GC^K}\{gc_i\}$ (3)

5.3 The overall decision system

In the final step of the proposed system, the facial expression analysis subsystem and the affective gesture analysis subsystem are integrated into a system which provides as result the possible emotions of the user, each accompanied by a degree of belief.

Although face consists the main "demonstrator" of user's emotion [9], the recognition of the accompanying gesture increases the confidence of the result of facial expression subsystem [2]. Further research is necessary to be carried out in order to define how powerful the influence of a gesture in the recognition of an emotion actually is. It would also be helpful to define which, face or gesture, is more useful for a specific application and change the impact of each subsystem on the final result.

In the current implementation the two subsystems are combined as a weighted sum: Let b_k be the degree of belief that the observed sequence presents the *k-th* emotional state, obtained from the facial expression analysis subsystem, and EI_k be the corresponding emotional state indicator, obtained from the affective gesture analysis subsystem, then the overall degree of belief d_k is given by:

$$d_k = w_1 \cdot b_k + w_2 \cdot EI_k \qquad (4)$$

where the weights w_1 and w_2 are used to account for the reliability of the two subsystems as far as the emotional state estimation is concerned. In this implementation we use w_1 =0.75 and w_2 =0.25. These values enables the affective gesture analysis subsystem to be important in cases where the facial expression analysis subsystem produces ambiguous results while at the same time leaves the latter subsystem to be the main contributing part in the overall decision system.

For the input sequence shown in Figure 1, the affective gesture analysis subsystem consistently provided a “surprise” selection. This was used to fortify the output of the facial analysis subsystem which was around 85%.

6. CONCLUSIONS – FUTURE WORK

In this paper we described a holistic approach to emotion modeling and analysis and their applications in MMI applications. 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. 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.

Future work in the affective modeling area, includes the enrichment of the gesture vocabulary with more affective gestures, as well as the relevant feature-based descriptions. With respect to the recognition part, more sophisticated methods of combination of detected expressions and gestures, mainly through a rule based system, are currently under investigation, along with algorithms that take into account general body posture information.

7. REFERENCES

[1] C. Wren, A. Azarbayejani, T. Darrel, and A. Pentland, “Pfinder: Real-time tracking of the human body,” *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 9, no. 7, pp. 780-785, 1997.

[2] D. McNeill, *Hand and mind: what gestures reveal about thought*, University of Chicago Press, Chicago, USA, 1992.

[3] G. Klir and B. Yuan, “*Fuzzy Sets and Fuzzy Logic, Theory and Application*”, Prentice Hall, New Jersey, 1995.

[4] J. Lin, Y. Wu, and T.S. Huang, “Modeling human hand constraints,” in *Proc. Workshop on Human Motion*, Dec. 2000, pp. 121-126.

[5] K. Scherer and P. Ekman, *Approaches to Emotion*, Lawrence Erlbaum Associates, 1984.

[6] K. Karpouzis, N. Tsapatsoulis and S. Kollias, “Moving to Continuous Facial Expression Space using the MPEG-4 Facial Definition Parameter (FDP) Set,” *in Proc. of SPIE Electronic Imaging 2000,* San Jose, CA, USA, January 2000.

[7] M. Davis and H. College, *Recognition of Facial Expressions*, Arno Press, New York, 1975.

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

[9] P. Ekman and W. Friesen, *The Facial Action Coding System*, Consulting Psychologists Press, San Francisco, CA, 1978.

[10] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz and J. Taylor, “Emotion Recognition in Human-Computer Interaction”, *IEEE Signal Processing Magazine*, January 2001.

[11] R. Kjeldsen and J. Kender, “Finding skin in color images,” in *Proc. 2nd Int. Conf. Automatic Face and Gesture Recognition*, 1996, pp. 312-317.

[12] R. Sharma, T.S. Huang, and V.I. Pavlovic, “A Multimodal Framework for Interacting With Virtual Environments,” *Human Interaction With Complex Systems,* C.A. Ntuen and E.H. Park, eds., pp. 53-71. Kluwer Academic Publishers, 1996.

[13] R. W. Picard, *Affective Computing*, MIT Press, Cambridge, MA, 2000.

[14] Y. Wu and T.S. Huang, “Hand modeling, analysis, and recognition for vision-based human computer interaction”, *IEEE Signal Processing Magazine*, vol. 18, iss. 3, pp. 51-60, May 2001.

[15] 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, Hindawi Publishing Corporation, October 2002.

[16] G. Votsis, A. Drosopoulos and S. Kollias, “A modular approach to facial feature segmentation on real sequences”, *Signal Processing, Image Communication,* vol. 18, pp. 67-89, 2003.

[17] L.R. Rabiner, “A tutorial on HMM and Selected Applications in Speech Recognition,” Proceedings of the IEEE, vol.77, no. 2, 1989.