Chapter V

Facial Expression and Gesture Analysis for Emotionally-Rich Man-Machine Interaction

Kostas Karpouzis, Amaryllis Raouzaiou, Athanasios Drosopoulos, Spiros Ioannou, Themis Balomenos, Nicolas Tsapatsoulis, and Stefanos Kollias
National Technical University of Athens, Greece

Abstract

This chapter presents a holistic approach to emotion modeling and analysis and their applications in Man-Machine Interaction applications. Beginning from a symbolic representation of human emotions found in this context, based on their expression via facial expressions and hand gestures, we show that it is possible to transform quantitative feature information from video sequences to an estimation of a user's emotional state. While these features can be used for simple representation purposes, in our approach they are utilized to provide feedback on the users' emotional state, hoping to provide next-generation interfaces that are able to recognize the emotional states of their users.

Introduction

Current information processing and visualization systems are capable of offering advanced and intuitive means of receiving input from 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 universal, in the sense that they are recognized among 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 of MMI applications (Picard, 2000).

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 the 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 individually examining these aspects of human expression, recent research (Cowie et al., 2001) has shown that even this approach can benefit from taking into account multimodal information. Consider a situation where the user sits in front of a camera-equipped computer and responds verbally to written or spoken messages from the computer: speech analysis can indicate periods of silence from the part of the user, thus informing the visual analysis module that it can use related data from the mouth region, which is essentially ineffective when the user speaks. Hand gestures and body pose provide another powerful means of communication. Sometimes, a simple hand action, such as placing one's hands over their ears, can pass on the message that they've had enough of what they are hearing more expressively than any spoken phrase.

In this chapter, we present a systematic approach to analyzing emotional cues from user facial expressions and hand gestures. In the Section "Affective analysis in MMI," we provide an overview of affective analysis of facial expressions and gestures, supported by psychological studies describing emotions as discrete points or areas of an "emotional space." The sections "Facial expression analysis" and "Gesture analysis" provide algorithms and experimen-

tal results from the analysis of facial expressions and hand gestures in video sequences. In the case of facial expressions, the motion of tracked feature points is translated to MPEG-4 FAPs, which describe their observed motion in a high-level manner. Regarding hand gestures, hand segments are located in a video sequence via color segmentation and motion estimation algorithms. The position of these segments is tracked to provide the hand's position over time and fed into a HMM architecture to provide affective gesture estimation.

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 the section "Multimodal affective analysis." The decision of the fuzzy system is based on rules obtained from the extracted features of actual video sequences showing emotional human discourse, as well as feature-based description of common knowledge of what everyday expressions and gestures mean. Results of the multimodal affective analysis system are provided here, while conclusions and future work concepts are included in the final section "Conclusions – Future work."

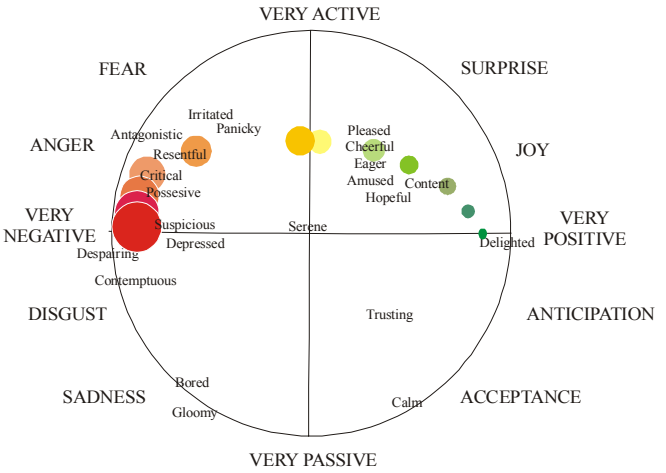
Effective Analysis in MMI

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 more closely matches our conception of what emotions are or, at least, how they are expressed and perceived.

Activation-emotion space (Whissel, 1989) is a representation that is both simple and capable of capturing a wide range of significant issues in emotion (Cowie et al., 2001). Perceived full-blown emotions are not evenly distributed in this space; instead, they tend to form a roughly circular pattern. From that and related evidence, Plutchik (1980) shows that there is a circular structure inherent in emotionality. In this framework, 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. A related extension is to think of primary or basic emotions as cardinal points on the periphery of an emotion

Figure 1. The Activation-emotion space



circle. Plutchik has offered a useful formulation of that idea, the "emotion wheel" (see Figure 1).

Activation-evaluation space is a surprisingly powerful device, which is increasingly being used in computationally oriented research. However, it has to be noted that such representations depend on collapsing the structured, high-dimensional space of possible emotional states into a homogeneous space of two dimensions. There is inevitably loss of information. Worse still, there are different ways of making the collapse lead to substantially different results. That is well illustrated in the fact that fear and anger are at opposite extremes in Plutchik's emotion wheel, but close together in Whissell's activation/emotion space. Thus, extreme care is needed to ensure that collapsed representations are used consistently.

MPEG-4 Based Representation

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. 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. Most of the techniques for facial animation are based on a well-known system for describing "all visually

distinguishable facial movements" called the Facial Action Coding System (FACS) (Ekman & Friesen, 1978). FACS is an anatomically oriented coding system, based on the definition of "Action Units" (AU) of a face that cause facial movements. An Action Unit could combine the movement of two muscles or work in the reverse way, i.e., split into several muscle movements. The FACS model has inspired the derivation of facial animation and definition parameters in the framework of MPEG-4 (Tekalp & Ostermann, 2000). In particular, the Facial Definition Parameter (FDP) and the Facial Animation Parameter (FAP) set were designed to allow the definition of a facial shape and texture. These sets eliminate 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.

Effective Facial Expression Analysis

There is a long history of interest in the problem of recognizing emotion from facial expressions (Ekman & Friesen, 1978), and extensive studies on face perception during the last 20 years (Davis & College, 1975). 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 (Karpouzis, Tsapatsoulis & Kollias, 2000).

Analysis of the emotional expression of a human face requires 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, 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.

Facial features can be viewed (Ekman & Friesen, 1975) as static (such as skin color), slowly varying (such as permanent wrinkles), or rapidly varying (such as raising the eyebrows) with respect to time evolution. Detection of the position and shape of the mouth, eyes and eyelids and extraction of related features are the targets of techniques applied to still images of humans. It has, however, been

shown (Bassili, 1979) that facial expressions can be more accurately recognized from image sequences, than from single still images. Bassili's experiments used point-light conditions, i.e., subjects viewed image sequences in which only white dots on a darkened surface of the face were visible. Expressions were recognized at above chance levels when based on image sequences, whereas only happiness and sadness were recognized when based on still images.

Effective Gesture Analysis

The detection and interpretation of hand gestures has become an important part of human computer interaction (MMI) in recent years (Wu & Huang, 2001). 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.

Gesture tracking and recognition

In general, human hand motion consists of the global hand motion and local finger motion. Hand motion capturing deals with finding the global and local motion of hand movements. Two types of cues are often used in the localization process: color cues (Kjeldsen & Kender, 1996) and motion cues (Freeman & Weissman, 1995). Alternatively, the fusion of color, motion and other cues, like speech or gaze, is used (Sharma, Huang & Pavlovic, 1996).

Hand localization is locating hand regions in image sequences. Skin color offers an effective and efficient way to fulfill this goal. According to the representation of color distribution in certain color spaces, current techniques of skin detection can be classified into two general approaches: nonparametric (Kjeldsen & Kender, 1996) and parametric (Wren, Azarbayejani, Darrel & Pentland, 1997).

To capture articulate hand motion in full degree of freedom, both global hand motion and local finger motion should be determined from video sequences. Different methods have been taken to approach this problem. One possible method is the appearance-based approach, in which 2-D deformable hand-shape templates are used to track a moving hand in 2-D (Darrell, Essa & Pentland, 1996). Another possible way is the 3-D model-based approach, which takes the advantages of *a priori* knowledge built in the 3-D models.

Meaningful gestures could be represented by both temporal hand movements and static hand postures. Hand postures express certain concepts through hand configurations, while temporal hand gestures represent certain actions by hand

movements. Sometimes, hand postures act as special transition states in temporal gestures and supply a cue to segment and recognize temporal hand gestures.

In certain applications, continuous gesture recognition is required and, as a result, the temporal aspect of gestures must be investigated. Some temporal gestures are specific or simple and could be captured by low-detail dynamic models. However, many high detail activities have to be represented by more complex gesture semantics, so modeling the low-level dynamics is insufficient. The HMM (Hidden Markov Model) technique (Bregler, 1997) and its variations (Darrell & Pentland, 1996) are often employed in modeling, learning, and recognition of temporal signals. Because many temporal gestures involve motion trajectories and hand postures, they are more complex than speech signals. Finding a suitable approach to model hand gestures is still an open research problem.

Facial Expression Analysis

Facial Features Relevant to Expression Analysis

Facial analysis includes a number of processing steps that attempt to detect or track the face, to locate characteristic facial regions such as eyes, mouth and nose, 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.

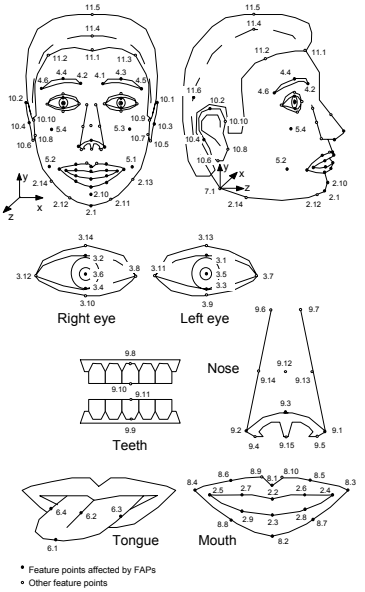
Although FAPs provide all the necessary elements for MPEG-4 compatible animation, they cannot be directly used for the analysis of expressions from video sequences, due to the absence of a clear quantitative definition framework. In order to measure FAPs in real image sequences, we have to define a mapping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

Table 1 provides the quantitative modeling of FAPs that we have implemented using the features labeled as f_i ($i=1..15$) (Karpouzis, Tsapatsoulis & Kollias, 2000). This feature set employs feature points that lie in the facial area and 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 2. Some of these points are constant during expressions and can be used as reference points. Distances between these points are used for normalization purposes (Raouzaiou, Tsapatsoulis, Karpouzis & Kollias, 2002).

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

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)	$f_{15}= D_{15\text{-}NEUTRAL} - D_{15}$

Figure 2. FDP feature points (adapted from (Tekalp & Ostermann, 2000))



Facial Feature Extraction

The facial feature extraction scheme used in the system proposed in this chapter is based on a hierarchical, robust scheme, coping with large variations in the appearance of diverse subjects, as well as the same subject in various instances within real video sequences (Votsis, Drosopoulos & Kollias, 2003). 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 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 with 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.

Experimental Results

Figure 3 below shows a characteristic frame from the “hands over the head” sequence. After skin detection and segmentation, the primary facial features are shown in Figure 4. Figure 5 shows the initial detected blobs, which include face and mouth. Figure 6 shows the estimates of the eyebrow and nose positions. Figure 7 shows the initial neutral image used to calculate the FP distances. In Figure 8 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 1.

Figure 3. A frame from the original sequence

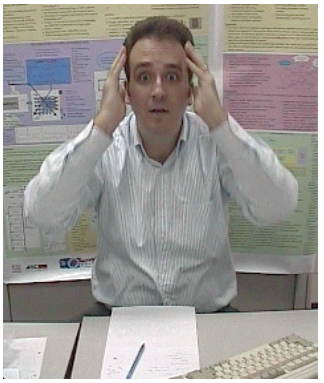


Figure 4. Detected primary facial features

Figure 5. The apex of anexpression

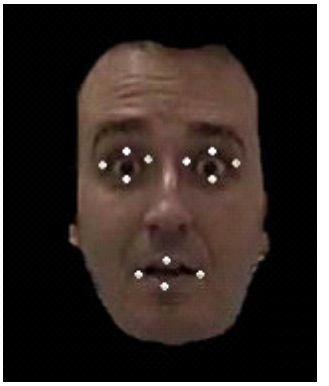


Figure 6. Detected facial features

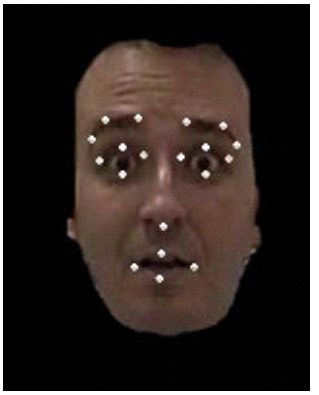


Figure 7. A neutral expression

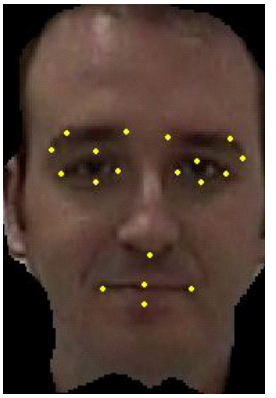
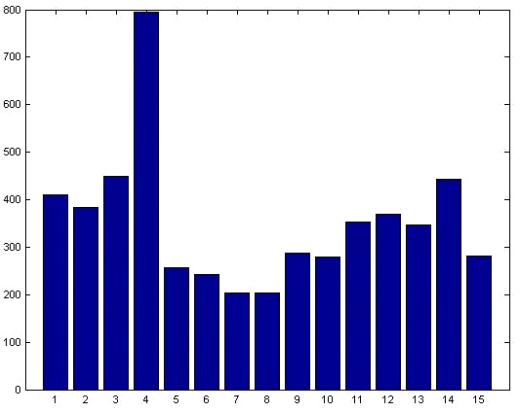


Figure 8. Estimated FAP values for Figure 6



Gesture Analysis

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 that 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 (Figure 9) 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 the 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 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 10). The skin color mask is then obtained from the skin probability matrix using thresholding (Figure 11). Possible moving areas are found by thresholding the difference pixels between the current frame and the next, resulting in the possible-motion mask (Figure 18). 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 11) and motion (Figure 18) masks

Figure 9. Abstract architecture of the hand tracking module

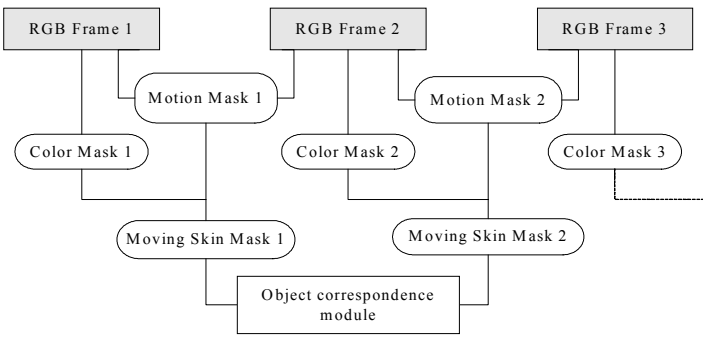


Figure 10. Skin Probability

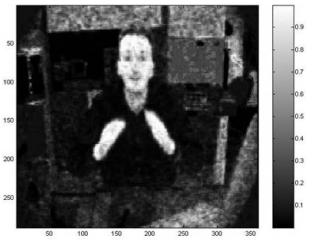


Figure 11. Thresholded skin probability (p> 0.8)

Figure 12. Distance transform of Figure 11

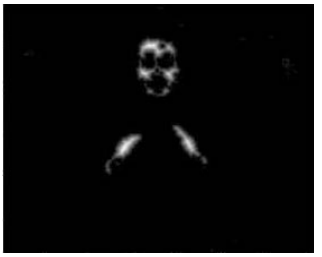


Figure 13. Markers extracted from Figure 12 (area smaller than 2% of the image)

Figure 14. Reconstruction of Figure 11 using Figure 13

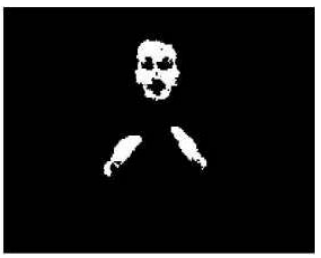


Figure 15. Closing of Figure 14, final color mask

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 (Figure 12) and only objects above the desired size are retained (Figure 13). 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 19. 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

Figure 16. Skin color probability for the input image

Figure 17. Initial color mask created with skin detection

Figure 18: Initial motion mask (after pixel difference thresholded to 10% of max.)

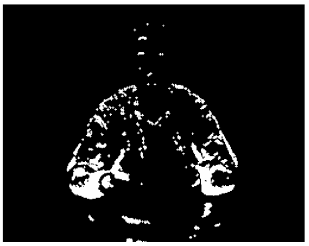


Figure 19. Moving hand segments after morphological reconstruction

Figure 20. Tracking of one hand object in the "lift of the hand" sequence

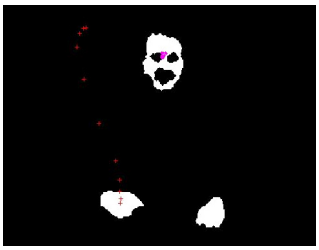
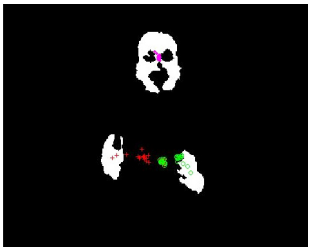


Figure 21. Tracking of both hand objects in the "clapping" sequence



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 20). 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 21).

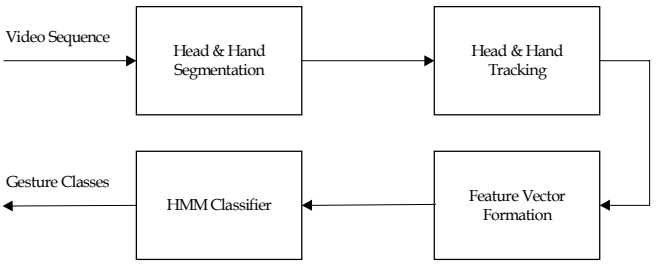
Following object matching in the subsequent moving skin masks, the mask flow is computed, i.e., a vector for each frame depicting the motion direction and magnitude of the frame's objects. The described algorithm is lightweight, allowing a rate of around 12 fps on a usual PC during our experiments, which is enough for continuous gesture tracking. The object correspondence heuristic makes it possible to individually track the hand segments correctly, at least during usual meaningful gesture sequences. In addition, the fusion of color and motion

information eliminates any background noise or artifacts, thus reinforcing the robustness of the proposed approach.

Gesture Classification Using HMMs

Figure 22 shows the architecture of the gesture classification subsystem. Head and hand segmentation and tracking have been described in previous sections, while the remaining blocks of this architecture are described in the following paragraphs.

Figure 22. A general framework for gesture classification through HMMs



The HMM Classifier

In Table 2 we present the utilized features that feed (as sequences of vectors) the HMM classifier, as well as the output classes of the HMM classifier.

Table 2: a) Features (inputs to HMM) and b) Gesture Classes (outputs of 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 $C_f=(X_f, Y_f)$ are the coordinates of the head centroid, $C_{rh}=(X_{rh}, Y_{rh})$ and $C_{lh}=(X_{lh}, Y_{lh})$ are the coordinates of the right and left hand centroids respectively
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

A general diagram of the HMM classifier is shown in Figure 23. 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 2. 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 $G_{3,3}$. More complex gesture classes, like the *hand clapping,* require as much as eight states to be efficiently modeled by the corresponding HMM. Some characteristics of our HMM implementation are presented below.

- The output probability for any state $G_{k,j}$ (*k* corresponds to the *id* of the HMM while *j* refers to the *id* of the state within a particular HMM) is obtained by a continuous probability density function (pdf). This choice has been made in order to decrease the amount of training data. In the discrete case, the size of the code book should be large enough to reduce quantization error and, therefore, a large amount of training data is needed to estimate the output probability. One problem with the continuous pdf is the proper selection of the initial values of density's parameters so as to avoid convergence in a local minimum.
- The output pdf of state $G_{k,j}$ is approximated using a multivariate normal distibution model, i.e.,

$$b_{k,j}(\mathbf{O}_i) = \frac{\exp\{-\frac{1}{2}(\mathbf{O}_i - \boldsymbol{\mu}_{k,j})^T \mathbf{C}_{k,j}^{-1}(\mathbf{O}_i - \boldsymbol{\mu}_{k,j})\}}{(2\pi)^{\frac{K}{2}} \cdot |\mathbf{C}_{k,j}|^{\frac{1}{2}}} \quad (1)$$

where $\mathbf{O}_i$ is *i-th* observation (input feature vector), $\boldsymbol{\mu}_{k,j}$ is the mean vector of state $G_{k,j}$, $\mathbf{C}_{k,j}$ is the respective covariance matrix and *K* is the number of components in $\mathbf{O}_i$ (in our case *K*=6). Initial values for $\boldsymbol{\mu}_{k,j}$ and $\mathbf{C}_{k,j}$ were obtained off-line by using statistical means. Re-estimation is executed using a variant of the Baum-Welch procedure to account for vectors (such as $\boldsymbol{\mu}_{k,j}$) and matrices (such as $\mathbf{C}_{k,j}$).

- Transition probabilities $a_{k,mn}$ between states $G_{k,m}$ and $G_{k,n}$ are computed by using the cumulative probability of $b_{k,m}(\mathbf{O}_i)$ gives the estimation of the transition probability, i.e., $a_{k,mn} = 1 - \Phi_{k,m}(\mathbf{O}_i)$. Note that, since the HMM is assumed to operate in a left-to-right mode, $a_{k,mn}=0, n<m$, $a_{k,mm}=1-a_{k,mn}$ at all times.

- The match score of feature vector sequence $\mathbf{O} = \mathbf{O}_1\mathbf{O}_2...\mathbf{O}_T$ given the model $\lambda_m(\mathbf{A}_m, \mathbf{B}_m, \boldsymbol{\pi}_m)$ (m=1,2,…,M) is calculated as follows:
  - We compute the best state sequence $\mathbf{Q}^*$ given the observation sequence $\mathbf{O}$, using Viterbi's algorithm, i.e.,

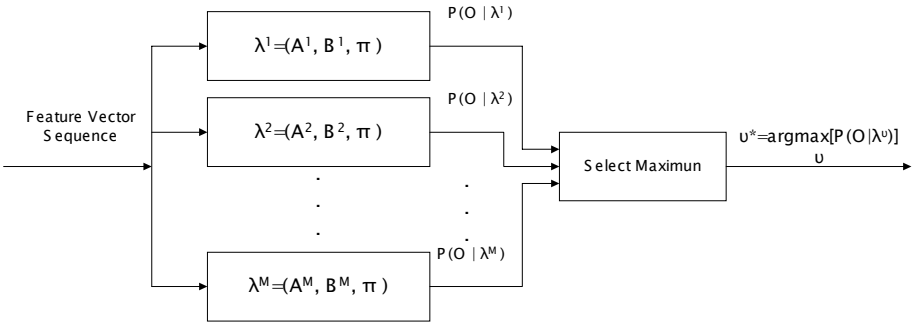
$$\mathbf{Q}^* = \arg\max_{\mathbf{Q}}\{P(\mathbf{Q}/\mathbf{O}, \lambda_m)\} \tag{2}$$

  - The match score of observation sequence $\mathbf{O}$ given the state sequence $\mathbf{Q}^*$ is the following quantity:

$$P^* = P(\mathbf{O}/\mathbf{Q}^*, \lambda_m) \tag{3}$$

It should be mentioned here that the final block of the architecture corresponds to a hard decision system, i.e., it selects the best-matched gesture class. However, when gesture classification is used to support the facial expression analysis process, the probabilities of the distinct HMMs should be used instead (soft decision system). In this case, since the HMMs work independently, their outputs do not sum up to one.

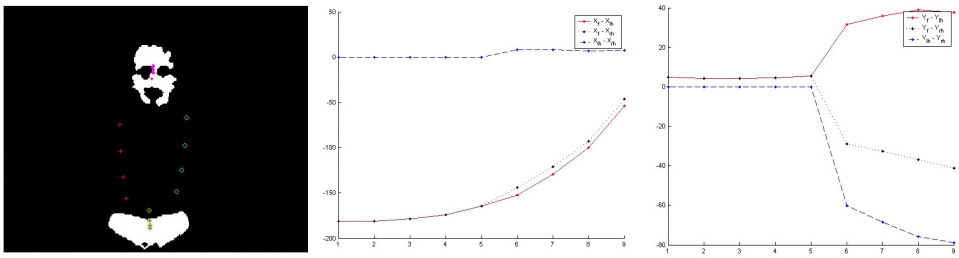
Figure 23. Block diagram of the HMM Classifier



Experimental Results

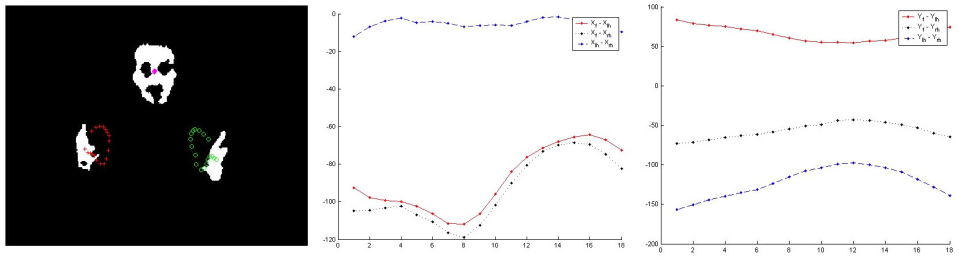
In the first part of our experiments, the efficiency of the features used to discriminate the various gesture classes is illustrated (Figure 24 to Figure 27). The first column shows a characteristic frame of each sequence and the tracked centroids of the head and left and right hand, while the remaining two columns show the evolution of the features described in the first row of Table 2, i.e., the difference of the horizontal and vertical coordinates of the head and hand segments. In the case of the first sequence, the gesture is easily discriminated since the vertical position of the hand segments almost matches that of the head, while in the closing frame of the sequence the three objects overlap. Overlapping is crucial to indicate that two objects are in contact during some point of the gesture, in order to separate this sequence from, e.g., the "lift of the hand" gesture. Likewise, during clapping, the distance between the two hand segments is zeroed periodically, with the length of the in-between time segments providing a measure of frequency, while during the "italianate" gesture the horizontal distance of the two hands follows a repetitive, sinusoidal pattern.

Figure 24. Hands over the head



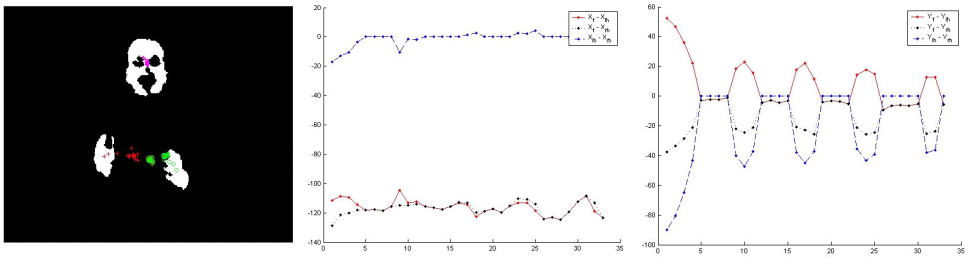
(a) *(b)* *(c)*

Figure 25. Italianate gesture



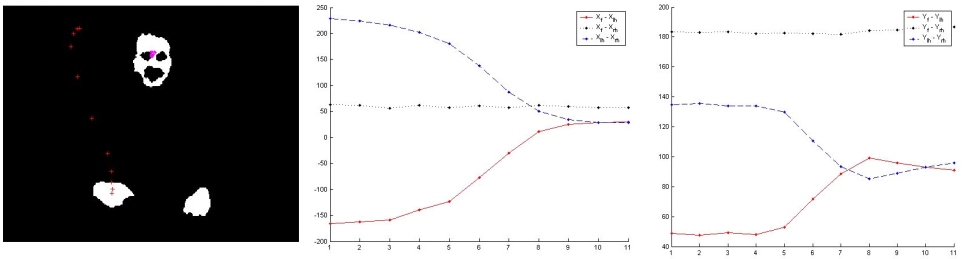
(a) *(b)* *(c)*

Figure 26. Hand clapping



(a) *(b)* *(c)*

Figure 27. Lift of the hand



(a)

Object centroids *crosses: left hand, circles: right hand, points: head*

(b)

Vertical object distances *dashes: X_{lh} - X_{rh},points: X_f- X_{rh}, line: X_f- X_{lh}Horizontal axis: framesVertical axis: pixels*

(c)

Horizontal object distances *dashes: Y_{lh}- Y_{rh},points: Y_f-Y_{rh}, line: Y_f-Y_{lh}Horizontal axis: framesVertical axis: pixels*

Experiments for testing the recognizing performance of the proposed algorithm were also 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 parameter 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 five frames per second so as to enable substantial motion to occur. An overall recognition

Table 3. Gesture classification results

Gesture Class	HC-LF	HC-HF	LH-LS	LH-HS	HoH-G	HoH-P	IG
Hand Clapping- Low Frequency (HC-LF)	5	0	0	0	0	0	0
Hand Clapping- High Frequency (HC-HF)	0	4	0	0	0	0	1
Lift of the Hand-Low Speed (LH-LS)	0	0	5	0	0	0	0
Lift of the Hand- High Speed (LH-HS)	0	0	0	5	0	0	0
Hands over the Head – Gesture (HoH-G)	0	0	0	0	5	0	0
Hands over the Head – Posture (HoH-P)	0	0	0	0	0	5	0
Italianate Gestures (IG)	0	1	0	0	0	0	4
Classification Rate (%)	100	80	100	100	100	100	80

rate of 94.3% was achieved. The experimental results are shown in the confusion matrix (Table 3).

From the results summarized in Table 3, we observe a mutual misclassification between "Italianate Gestures" (IG) and "Hand Clapping – High Frequency" (HC - HF). 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.

Multimodal Effective Analysis

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 sub-system a 15-element length feature vector $\underline{f}$ that corresponds to the 15 features f_i shown in Table 1. 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 correspondingly involved FAPs.

Expression profiles are also used to capture variations of FAPs (Raouzaiou, Tsapatsoulis, Karpouzis & Kollias, 2002). For example, the range of variations of FAPs for the expression "surprise" is shown in Table 4.

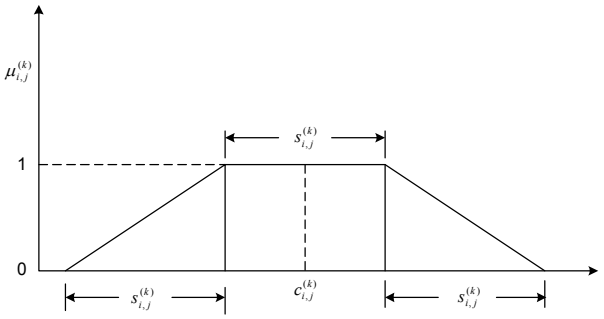
Let $X_{i,j}^{(k)}$ be the range of variation of FAP F_j involved in the *k-th* profile $P_i^{(k)}$ of emotion *i*. If $c_{i,j}^{(k)}$ and $s_{i,j}^{(k)}$ are the middle point and length of interval $X_{i,j}^{(k)}$ respectively, then we describe a fuzzy class $A_{i,j}^{(k)}$ for F_j, using the membership function $\mu_{i,j}^{(k)}$ shown in Figure 28. 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:

$$p_i^{(k)} = \prod_{A_{i,j}^{(k)} \in \Delta_{i,j}^{(k)}} r_{i,j}^{(k)} \quad \text{and} \quad b_i = \max_k (p_i^{(k)}), \tag{4}$$

Table 4. Profiles for the archetypal emotion surprise

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

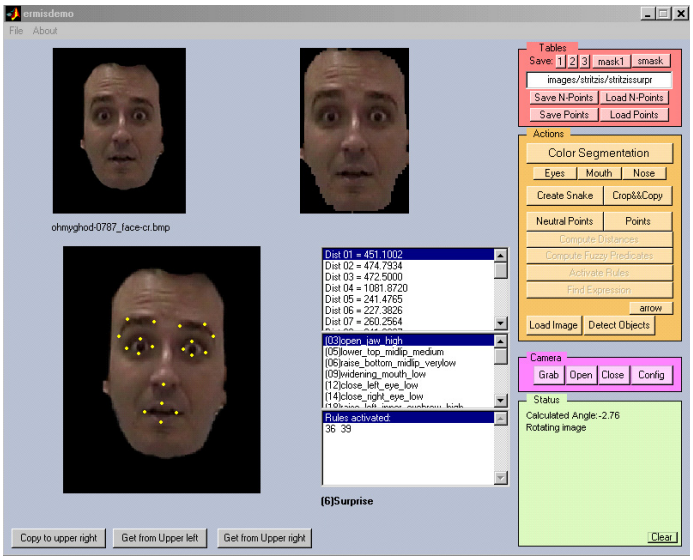
Figure 28. The form of membership functions



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 (Klir & Yuan, 1995).

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 $u(a,b)=\max(a,b)$.

Figure 29. Facial expression analysis interface



An efficient implementation of the emotion analysis system has been developed in the framework of the IST ERMIS project (www.image.ntua.gr/ermis). In the system interface shown in Figure 29, one can observe an example of the calculated FP distances, the profiles selected by the facial expression analysis subsystem and the recognized emotion ("surprise").

Effective 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 the section "Gesture analysis," 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 a particular gesture class. Table 5 below shows the correlation between some detectable gestures with the six archetypal expressions.

Given a particular context of interaction, gesture classes corresponding to the same emotional are combined in a "logical OR" form. Table 5 shows that a particular gesture may correspond to more than one gesture class carrying

Table 5. Correlation between gestures and emotional states

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*

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*.

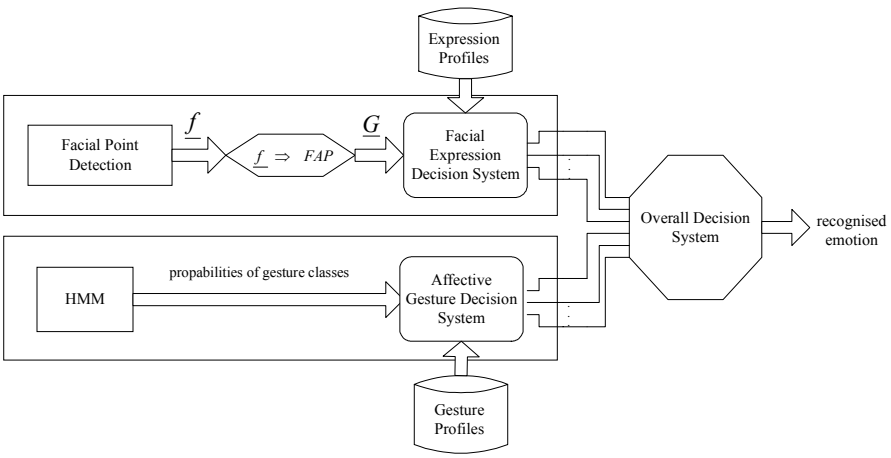
In practice, the gesture class probabilities derived by the HMM classifier are transformed to emotional state indicators by using the information of Table 5. Let EI_k be the emotional indicator of emotional state k ($k \in \{1,2,3,4,5,6\}$ corresponds to one of the emotional states presented in Table 5 in the order of appearance, i.e., 1->Joy, 6->Surprise), $GCS = \{gc_1, gc_2, \ldots, gc_N\}$ 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:

$$EI_k = \max_{gc_i \in GC^K} \{gc_i\} \tag{5}$$

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, as shown in Figure 30, into a system which provides as a result the possible emotions of the user, each accompanied by a degree of belief.

Figure 30. Block diagram of the proposed scheme



Although face is considered the main "demonstrator" of user's emotion (Ekman & Friesen, 1975), the recognition of the accompanying gesture increases the confidence of the result of the facial expression subsystem. 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:

$$d_k = w_1 \cdot b_k + w_2 \cdot EI_k \tag{6}$$

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 enable the affective gesture analysis subsystem to be important in cases where the facial expression analysis subsystem produces ambiguous results, while at the same time leave the latter subsystem to be the main contributing part in the overall decision system.

For the input sequence shown in Figure 3, 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%.

Conclusions – Future Work

In this chapter, 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 and hand gestures, 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 and hand movement. 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 user's emotional state while in front of a computer.

Future work in the affective modeling area includes the enrichment of the gesture vocabulary with more affective gestures and feature-based descrip-

tions. 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.

References

Bassili, J. N. (1979). Emotion recognition: The role of facial movement and the relative importance of upper and lower areas of the face. *Journal of Personality and Social Psychology*, 37, 2049-2059.

Bregler, C. (1997). Learning and recognition human dynamics in video sequences. *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 568-574.

Cowie, R. et al. (2001). Emotion Recognition in Human-Computer Interaction. *IEEE Signal Processing Magazine,* 1, 32-80.

Darrell, T. & Pentland, A. (1996). Active gesture recognition using partially observable Markov decision processes. In *Proc. IEEE Int. Conf. Pattern Recognition,* 3, 984-988.

Darrell, T., Essa, I. & Pentland, A. (1996). Task-Specific Gesture Analysis in Real-Time Using Interpolated Views. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 18(12), 1,236-1,242.

Davis, M. & College, H. (1975). Recognition of Facial Expressions. New York: Arno Press.

Ekman, P. & Friesen, W. (1975). Unmasking the Face. New York: Prentice-Hall.

Ekman, P. & Friesen, W. (1978). The Facial Action Coding System. San Francisco, CA: Consulting Psychologists Press.

Faigin, G. (1990). The Artist's Complete Guide to Facial Expressions. New York: Watson-Guptill.

Freeman, W. T. & Weissman, C. D. (1995). Television Control by Hand Gestures. *Proc. Int'l Workshop on Automatic Face and Gesture Recognition*, Switzerland, 179-183.

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

Kjeldsen, R. & Kender, J. (1996). Finding skin in color images. *Proc. 2nd Int. Conf. Automatic Face and Gesture Recognition*, 312-317.

Klir, G. & Yuan, B. (1995). Fuzzy Sets and Fuzzy Logic, Theory and Application. New Jersey: Prentice-Hall.

Picard, R. W. (2000). Affective Computing. Cambridge, MA: MIT Press.

Plutchik, R. (1980). Emotion: A psychoevolutionary synthesis. New York: Harper and Row.

Raouzaiou, A., Tsapatsoulis, N., Karpouzis, K. & Kollias, S. (2002). Parameterized facial expression synthesis based on MPEG-4. *EURASIP Journal on Applied Signal Processing*, 10, 1021-1038.

Sharma, R., Huang, T. S. & Pavlovic, V. I. (1996). A Multimodal Framework for Interacting with Virtual Environments. In C. A. Ntuen and E. H. Park (Eds), *Human Interaction With Complex Systems*. Kluwer Academic Publishers.

Tekalp, M. & Ostermann, J. (2000). Face and 2-D mesh animation in MPEG-4. *Image Communication Journal*, 15(4-5), 387-421.

Votsis, G., Drosopoulos, A. & Kollias, S. (2003). A modular approach to facial feature segmentation on real sequences. *Signal Processing, Image Communication,* 18, 67-89.

Whissel, C. M. (1989). The dictionary of affect in language. In R. Plutchnik & H. Kellerman (Eds), *Emotion: Theory, research and experience: volume 4, The measurement of emotions*. New York: Academic Press.

Wren, C., Azarbayejani, A., Darrel, T. & Pentland, A. (1997). Pfinder: Real-time tracking of the human body. *IEEE Trans. Pattern Anal. Machine Intell.*, 9(7), 780-785.

Wu, Y. & Huang, T. S. (2001). Hand modeling, analysis, and recognition for vision-based human computer interaction. *IEEE Signal Processing Magazine*, 18(3), 51-60.