Emotion Representation for Virtual Environments

K. Karpouzis, A. Raouzaiou, N. Tsapatsoulis and S. Kollias
Image, Video and Multimedia Systems Laboratory
Department of Electrical and Computer Engineering,
National Technical University of Athens
15780, Zographou, Athens, Greece
{kkarpou, araouz, ntsap}@image.ece.ntua.gr, stefanos@cs.ntua.gr

Abstract— **Research on networked applications 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. Human faces may act as visual interfaces that help users feel at home when interacting with a computer because they are accepted as the most expressive means for communicating and recognizing emotions. Thus, a lifelike human face can enhance interactive applications by providing straightforward feedback to and from the users and stimulating emotional responses from them. Thus, virtual environments can employ believable, expressive characters since such features significantly enhance the atmosphere of a virtual world and communicate messages far more vividly than any textual or speech information. In this paper, we present an abstract means of description of facial expressions, by utilizing concepts included in the MPEG-4 standard. Furthermore, we exploit these concepts to synthesize a wide variety of expressions using a reduced representation, suitable for networked and lightweight applications.**

Keywords-emotional representation; MPEG-4; networked virtual environments; avatars; expression synthesis

I. 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 [19] of MMI applications.

From the synthesis point of view, the ability to simulate lifelike interactive characters has many applications. A human face may act as visual interface that helps users feel at home when interacting with a computer. Such an agent can play the role of a personal assistant or troubleshooter, a tutor or even a Sensitive Artificial Listener that engages in discussions by providing dialogue cues based on the content of the actual responses from its users. In general, faces make good interface elements since they are accepted as the most expressive means for communicating, as well as recognizing emotions. Thus, a lifelike human face can enhance interactive applications such as information booths, e-commerce front-ends or educational systems by providing straightforward feedback to and from the users and stimulating emotional responses from them. Besides this, the gaming and entertainment industries can benefit from employing believable, expressive characters since such features significantly enhance the atmosphere of a virtual world and communicate messages far more vividly than any textual or speech information [7].

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 [17] 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. Inversely, the same verbal response from the part of the user, e.g. the phrase "what do you think", can be interpreted in a different manner when pronunciation or facial expression are also taken into account and indicate question, hopelessness or even irony.

7th International Conference on Telecommunications - ConTEL 2003
ISBN: 953-184-052-0, June 11-13, 2003, Zagreb, Croatia

Hand gestures and body pose provide another powerful means of communication. Sometimes, a simple hand action, such as placing ones' 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. Besides conveying that message to an analysis system, gestures and pose can also provide benefit to multiuser environments, where communication is traditionally reduced to text or text-to-speech voice.

Multiuser environments are an obvious testbed of emotionally rich MMI systems that utilize results from both analysis and synthesis notions. Simple chat applications can be transformed into powerful chat rooms, where different users interact, with or without the presence of avatars that take part in this process, taking into account the perceived expressions of the users. Also, virtual malls or museums can benefit from analyzing user responses in terms of how they feel about the system itself or the choices offered and adapt themselves to provide a more satisfying experience. The adoption of token-based animation in the MPEG-4 framework benefits such networked applications, since the communication of simple, symbolic parameters is, in this context, enough to analyze, as well as synthesize facial expression, hand gestures and body motion. While current applications take little advantage from this technology, research results show that its powerful features will reach the consumer level in a short period of time.

In this paper, we present an integrated approach to analyzing emotional cues from user facial expressions and hand gestures. In section II we provide results from psychological studies that describe emotions as discrete points or areas of an "emotional space"; this is essential in order to describe them using high level symbols, such as facial feature movement. Sections III and IV provide algorithms and results from the analysis of facial expressions and hand gestures in video sequences. These modals are treated in a different manner, since the tracked features are inherently diverse. More specifically, facial features are located in a neutral expression and then tracked throughout the discourse; the measured distance from their neutral position is translated to MPEG-4 compatible FAPs, which describe their observed motion in a higher-level manner. Regarding hand gestures, hand segments are located in a video sequence via color segmentation algorithms and then tracked to provide the hand's position over time. These findings are combined with kinematic constraints of the upper body to provide body posture information which is utilized both to recognize specific gestures, as well as provide low level information concerning hand movement, such as direction and speed of movement, frequency of repeating gestures, etc. Again, the observed or deduced body posture is described using MPEG-4 BAPs and BBA (*Bone Based Animation*) information, which is essential to transform the information from the video signal to symbolic tokens.

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 V. The decision of the fuzzy system is based on rules obtained from the extracted features of actual images and 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 facial expression and hand gesture analysis subsystems are provided, along with concepts of the application of the above findings to affective synthetic characters.

II. EMOTIONAL GESTURES IN MMI

A. *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 [17] 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 level*: Research has recognised 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.

The axes of the activation-evaluation space reflect those themes. The vertical axis shows activation level, the horizontal axis evaluation. A basic attraction of that arrangement is that it provides a way of describing emotional states which is more tractable than using words, but which can be translated into and out of verbal descriptions. Translation is possible because emotion-related words can be understood, at least to a first approximation, as referring to positions in activation-emotion space. Various techniques lead to that conclusion, including factor analysis, direct scaling, and others [21].

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, [18] 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

ConTEL 2003, ISBN: 953-184-052-0

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 has offered a useful formulation of that idea, the 'emotion wheel' (see **Figure 1**).

Activation-evaluation space is a surprisingly powerful device, and it has been increasingly used in computationally oriented research. However, it has to be emphasized that representations of that kind 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; and worse still, 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. Extreme care is, thus, needed to ensure that collapsed representations are used consistently.

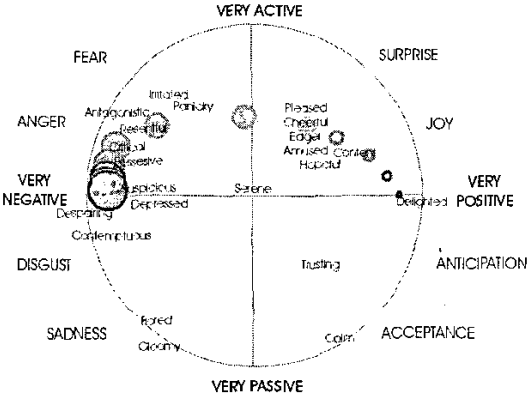

Figure 1: The Activation-emotion space

B. MPEG-4 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. The goal of FBA definition is the animation of both realistic and cartoonist characters. Thus, MPEG-4 has defined a large set of parameters and the user can select subsets of these parameters according to the application. The FBA part can be also combined with multimodal input (e.g. linguistic and paralinguistic speech analysis).

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). 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 movement. The FACS model has inspired the derivation of facial animation and definition parameters in the framework of the ISO MPEG-4 standard [14]. 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. These results indicate that facial expressions, possibly combined with gestures and speech, when the latter is available, provide cues that can be used to perceive a person's emotional state.

The second version of the standard, following the same procedure with the facial definition and animation (through FDPs and FAPs), describes the anatomy of the human body with groups of distinct tokens, eliminating the need to specify the topology of the underlying geometry. These tokens can then be mapped to automatically detected measurements and indications of motion on a video sequence, thus, they can help to estimate a real motion conveyed by the subject and, if required, approximate it by means of a synthetic one.

In general, an MPEG body is a collection of nodes. The Body Definition Parameter (BDP) set provides information about body surface, body dimensions and texture, while Body Animation Parameters (BAPs) transform the posture of the body. BDPs describe the topology of the human skeleton, taking into consideration joints limitations and independent degrees of freedom in the skeleton model of the different body parts.

C. BBA (Bone Based Animation)

The MPEG-4 BBA offers a standardized interchange format extending the MPEG-4 FBA [13]. In BBA the skeleton is a hierarchical structure made of bones. In this hierarchy every bone has one parent and can have as children other bones, muscles or 3D objects. For the movement of every bone we have to define the influence of this movement to the skin of our model, the movement of its children and the related inverse kinematics.

In the BBA stream the rotation is represented as Euler angles. It contains all the animation frames or data at the temporal key frames, where decoder will compute the intermediate frames by temporal interpolation (*linear* interpolation for *translation* and *scale* and *spherical linear quaternion* interpolation for *rotation* and *scaleOrientation*).

Bone based representations benefit both the synthesis and the analysis of hand gestures, since they are closer to human conceptions of body posture than mere motion information. Knowledge of the kinematic structure and constraints of the upper body, as well as the limited mobility of the torso in MMI applications make BAP and BBA representations interchangeable, for animation purposes [20].

III. FACIAL EXPRESSIONS

There is a long history of interest in the problem of recognizing emotion from facial expressions [15], and extensive studies on face perception during the last twenty years [12]. The salient issues in emotion recognition from faces are parallel in some respects to the issues associated with voices, but divergent in others.

As in speech, a long established tradition attempts to define the facial expression of emotion in terms of qualitative targets, i.e. static positions capable of being displayed in a still photograph. The still image usually captures the apex of the expression, i.e. the instant at which the indicators of emotion are most marked. More recently emphasis, has switched towards descriptions that emphasize gestures, i.e. significant movements of facial features.

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 [11].

D. Facial expressions

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 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.

Facial features can be viewed [17] as either static (such as skin color), or 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, particularly eyelids, wrinkles and extraction of features related to them are the targets of techniques applied to still images of humans. It has, however, been shown [9], that facial expressions can be more accurately recognized from image sequences, than from a single still image. His 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 at above chance levels when based on still images. Techniques which attempt to identify facial gestures for emotional expression characterization face the problems of locating or extracting the facial regions or features, computing the spatio-temporal motion of the face through optical flow estimation, and introducing geometric or physical muscle models describing the facial structure or gestures.

IV. VISUAL GESTURE INTERPRETATION

The detection and interpretation of hand gestures has become an important part of human computer interaction (MMI) in recent years. 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. To benefit from the use of gestures in MMI it is necessary to provide the means by which they can be interpreted by computers. The MMI interpretation of gestures requires that dynamic and/or static configurations of the human hand, arm, and even other parts of the human body, be measurable by the machine. First attempts to address this problem resulted in mechanical devices that directly measure hand and/or arm joint angles and spatial position. The so-called glove-based devices best represent this solutions' group.

Since the processing of visual information provides strong cues in order to infer the states of a moving object through time, vision-based techniques provide at least adequate, alternatives to capture and interpret human hand motion. At the same time, applications can benefit from the fact that vision systems can be very cost efficient and do not affect the natural interaction with the user. These facts serve as the motivating forces for research in the modeling, analysis, animation, and recognition of hand gestures. 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. 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, 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. Evaluation of a particular gesture recognition approach encompasses accuracy, robustness, and speed, as well as the variability in the number of different classes of hand/arm movements it covers.

ConTEL 2003, ISBN: 953-184-052-0

E. Gesture Modeling

In order to systematically discuss the literature on gesture interpretation, it is important to first consider what model the authors have used for the hand gesture. In fact, the scope of a gestural interface for MMI is directly related to the proper modeling of hand gestures. The modeling of hand gestures depends primarily on the intended application within the MMI context. For a given application, a very coarse and simple model may be sufficient. However, if the purpose is a natural-like interaction, a model has to be established that allows many, if not all, natural gestures to be interpreted by the computer.

Human hand motion is highly articulate, because the hand consists of many connected parts that lead to complex kinematics. At the same time, hand motion is also highly constrained, which makes it difficult to model. Usually, the hand can be modeled in several aspects such as shape [10], kinematical structure [8], dynamics [5], and semantics.

F. Gesture Recognition

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. Although hand gestures are complicated to model because the meanings of hand gestures depend on people and cultures, a set of specific hand gesture vocabulary can always be predefined in many applications, such as virtual environment (VE) applications, so that the ambiguity can be limited.

Different from sign languages, the gesture vocabulary in VE applications is structured and disambiguated. Some simple controlling, commanding, and manipulative gestures are defined to fulfill natural interaction such as pointing, navigating, moving, rotating, stopping, starting, and selecting. These gesture commands can be simple in the sense of motion; however, many different hand postures are used to differentiate and switch among the commanding modes. For example, only if we know a gesture is a pointing gesture would it make sense to estimate its pointing direction. View-independent hand posture recognition is a natural requirement in many VE applications. In most cases, because users do not know where the cameras are, the naturalness and immersiveness will be ruined if users are obliged to issue commands to an unknown direction.

G. Hand detection and tracking for MMI

Gesture analysis research follows two different approaches that work in parallel. The first approach treats a hand gesture as a two- or three dimensional signal that is communicated via hand movement from the part of the user; as a result, the whole analysis process merely tries to locate and track that movement, so as to recreate it on an avatar or translate it to specific, predefined input interface, e.g. raising hands to draw attention or indicate presence in a virtual classroom.

The low level results of the approach can be extended, taking into account that hand gestures are a powerful expressive means. The expected result is to understand gestural interaction as a higher-level feature and encapsulate it into an original modal, complementing speech and image analysis in an affective MMI system [2]. This transformation of a gesture from a time-varying signal into a symbolic level helps overcome problems such as the proliferation of available gesture representations or failure to notice common features in them. In general, one can classify hand movements with respect to their function as:

- *Semiotic*: these gestures are used to communicate meaningful information or indications
- *Ergotic*: manipulative gestures that are usually associated with a particular instrument or job and
- *Epistemic*: again related to specific objects, but also to the reception of tactile feedback.

Semiotic hand gestures are considered to be connected, or even complementary, to speech in order to convey a concept or emotion. Especially two major subcategories, namely *deictic gestures* and *beats*, i.e. gestures that consist of two discrete phases, are usually semantically related to the spoken content and used to emphasize or clarify it [3]. This relation is also taken into account in [1] and provides a positioning of gestures along a continuous space.

In this study, certain gestures are considered spontaneous, free form movements of the hands during speech (gesticulation), while others, termed *emblems*, are indicative of a specific emotion or action, such as an insult. An interesting conclusion in [3] is that the alternative use of gestures and speech in order to comprehend the communicated emotion or idea makes the whole concept of body language obsolete. Indeed, the study shows that instead of being "mere embellishments" of spoken content, gestures possess a number of para-linguistic properties. For example, such gestures convey a specific meaning only when considered as a whole, not as mere collections of low level hand movements. While spoken words are usually unambiguous and can be semantically interpreted only when in a complete sentence or paragraph, gestures are atomic when it comes to conveying an idea and typically their actual form depends on the personality and current emotional state of a specific speaker. As a result, gestures cannot be analyzed with the same tools used to process the other modals of human discourse. In the case of gesticulation, we can regard gestures as functions of hand movement over time; the result of this approach is that the quantitative values of this representation, such as speed, direction or repetition, can be associated to emotion-related values, such as activation. This essentially means that in many cases we do not need to recognize specific gestures to deduce information about the users' emotional state, but merely track the movement of their arms through time. This concept can also help us distinguish a specific gesture from a collection of similar hand movements: for example, the "raise hand" gesture in a classroom or discussion and the "go away" or "I've had enough" gestures are similar when it comes to hand movement, since in both cases the hand is raised vertically. The only way to differentiate them is to compare the speed of the upward movement in both cases: in the latter case the

hand is raised in a much more abrupt manner. In our approach, such feedback is invaluable, since we try to analyze the users' emotional state by taking into account a combination of both gesture- and face-related features and not decide based on merely one of the two modals.

V. FROM FEATURES TO SYMBOLS

In order to estimate the users' emotional state in a MMI context, we must first describe the six archetypal expressions in a symbolic manner, using easily and robustly estimated tokens. FAPs and BAPs or BBA representations make good candidates for describing quantitative facial and hand motion features. The use of these parameters serves several purposes such as compatibility of created synthetic sequences with the MPEG-4 standard and increase of the range of the described emotions – archetypal expressions occur rather infrequently and in most cases emotions are expressed through variation of a few discrete facial features related with particular FAPs.

Based on elements from psychological studies [16], [4], [6], we have described the six archetypal expressions using MPEG-4 FAPs, which is illustrated in Table 1. In general, these expressions can be uniformly recognized across cultures and are therefore invaluable in trying to analyze the users' emotional state.

Joy	*open_jaw(F_3), lower_t_midlip(F_4), raise_b_midlip(F_5), stretch_l_cornerlip(F_6), stretch_r_cornerlip(F_7), raise_l_cornerlip(F_{12}), raise_r_cornerlip(F_{13}), close_t_l_eyelid(F_{19}), close_t_r_eyelid(F_{20}) , close_b_l_eyelid(F_{21}), close_b_r_eyelid(F_{22}), raise_l_m_eyebrow (F_{33}), raise_r_m_eyebrow(F_{34}), lift_l_cheek (F_{41}), lift_r_cheek(F_{42}), stretch_l_cornerlip_o (F_{53}), stretch_r_cornerlip_o(F_{54})*
Sadness	*close_t_l_eyelid(F_{19}), close_t_r_eyelid(F_{20}), close_b_l_eyelid(F_{21}), close_b_r_eyelid(F_{22}), raise_l_i_eyebrow(F_{31}), raise_r_i_eyebrow (F_{32}), raise_l_m_eyebrow(F_{33}), raise_r_m_eyebrow(F_{34}), raise_l_o_eyebrow (F_{35}), raise_r_o_eyebrow(F_{36})*
Anger	*lower_t_midlip(F_4), raise_b_midlip(F_5), push_b_lip(F_{16}), depress_chin(F_{18}), close_t_l_eyelid(F_{19}), close_t_r_eyelid(F_{20}), close_b_l_eyelid(F_{21}), close_b_r_eyelid(F_{22}), raise_l_i_eyebrow(F_{31}), raise_r_i_eyebrow (F_{32}), raise_l_m_eyebrow(F_{33}), raise_r_m_eyebrow(F_{34}), raise_l_o_eyebrow (F_{35}), raise_r_o_eyebrow(F_{36}), squeeze_l_eyebrow(F_{37}), squeeze_r_eyebrow (F_{38})*
Fear	*open_jaw(F_3), lower_t_midlip(F_4), raise_b_midlip(F_5), lower_t_lip_lm(F_8), lower_t_lip_rm(F_9), raise_b_lip_lm (F_{10}), raise_b_lip_rm(F_{11}), close_t_l_eyelid (F_{19}), close_t_r_eyelid(F_{20}), close_b_l_eyelid (F_{21}), close_b_r_eyelid(F_{22}), raise_l_i_eyebrow (F_{31}), raise_r_i_eyebrow(F_{32}), raise_l_m_eyebrow(F_{33}), raise_r_m_eyebrow (F_{34}), raise_l_o_eyebrow(F_{35}), raise_r_o_eyebrow (F_{36}), squeeze_l_eyebrow (F_{37}), squeeze_r_eyebrow (F_{38})*
Disgust	*open_jaw (F_3), lower_t_midlip (F_4), raise_b_midlip (F_5), lower_t_lip_lm (F_8), lower_t_lip_rm (F_9), raise_b_lip_lm (F_{10}), raise_b_lip_rm (F_{11}), close_t_l_eyelid (F_{19}), close_t_r_eyelid (F_{20}), close_b_l_eyelid (F_{21}), close_b_r_eyelid(F_{22}), raise_l_m_eyebrow (F_{33}), raise_r_m_eyebrow(F_{34}), lower_t_lip_lm_o (F_{55}), lower_t_lip_rm_o (F_{56}), raise_b_lip_lm_o (F_{57}), raise_b_lip_rm_o (F_{58}), raise_l_cornerlip_o (F_{59}), raise_r_cornerlip_o (F_{60})*
Surprise	*open_jaw (F_3), raise_b_midlip (F_5), stretch_l_cornerlip (F_6) stretch_r_cornerlip (F_7), raise_b_lip_lm(F_{10}), raise_b_lip_rm(F_{11}), close_t_l_eyelid (F_{19}), close_t_r_eyelid (F_{20}), close_b_l_eyelid (F_{21}), close_b_r_eyelid (F_{22}), raise_l_i_eyebrow(F_{31}), raise_r_i_eyebrow (F_{32}), raise_l_m_eyebrow (F_{33}), raise_r_m_eyebrow (F_{34}), raise_l_o_eyebrow (F_{35}), raise_r_o_eyebrow (F_{36}), squeeze_l_eyebrow (F_{37}), squeeze_r_eyebrow (F_{38}), stretch_l_cornerlip_o (F_{53}), stretch_r_cornerlip_o (F_{54})*

Table 1: *FAPs vocabulary for archetypal expression description*

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. In order to measure FAPs in real image sequences, we define a mapping between them and the movement of specific FDP feature points (FPs), which correspond to salient points on the human face. This quantitative description of FAPs provides the means of bridging the gap between expression analysis and synthesis. In the expression analysis case, the non-additive property of the FAPs can be addressed by a fuzzy rule system.

Quantitative modeling of FAPs is implemented using the features labeled as f_i ($i=1..15$) in Table 2 [11]. The feature set employs feature points that lie in the facial area and, in the controlled environment of MMI applications, can be automatically detected and tracked. It consists of distances, noted as $s(x,y)$, where x and y correspond to Feature Points [14], between these protuberant points, some of which are constant during expressions and are used as reference points; distances between these reference points are used for normalization purposes [22]. The units for f_i are identical to those corresponding to FAPs, even in cases where no one-to-one relation exists.

FAP name	Feature description for the	Utilized feature	Unit
Squeeze_l_eyebrow (F_{37})	$D_1=s(4.5,3.11)$	$f_1=D_{1\text{-}NEUTRAL}-D_1$	*ES*
Squeeze_r_eyebrow (F_{38})	$D_2=s(4.6,3.8)$	$f_2=D_{2\text{-}NEUTRAL}-D_2$	*ES*
Lower_t_midlip (F_4)	$D_3=s(9.3,8.1)$	$f_3=D_3-D_{3\text{-}NEUTRAL}$	*MNS*
Raise_b_midlip (F_5)	$D_4=s(9.3,8.2)$	$f_4=D_{4\text{-}NEUTRAL}-D_4$	*MNS*
Raise_l_l_eyebrow (F_{31})	$D_5=s(4.1,3.11)$	$f_5=D_5-D_{5\text{-}NEUTRAL}$	*ENS*
Raise_r_l_eyebrow (F_{32})	$D_6=s(4.2,3.8)$	$f_6=D_6-D_{6\text{-}NEUTRAL}$	*ENS*
Raise_l_o_eyebrow (F_{35})	$D_7=s(4.5,3.7)$	$f_7=D_7-D_{7\text{-}NEUTRAL}$	*ENS*

ConTEL 2003, ISBN: 953-184-052-0

Raise_r_o_eyebrow (F_{36})	$D_8=s(4.6,3.12)$	$f_8=D_8-D_{8\text{-}NEUTRAL}$	*ENS*
Raise_l_m_eyebrow (F_{33})	$D_9=s(4.3,3.7)$	$f_9=D_9-D_{9\text{-}NEUTRAL}$	*ENS*
Raise_r_m_eyebrow (F_{34})	$D_{10}=s(4.4,3.12)$	$f_{10}=D_{10}-D_{10\text{-}NEUTRAL}$	*ENS*
Open_jaw (F_3)	$D_{11}=s(8.1,8.2)$	$f_{11}=D_{11}-D_{11\text{-}NEUTRAL}$	*MNS*
close_t_l_eyelid (F_{19}) – *close_b_l_eyelid* (F_{21})	$D_{12}=s(3.1,3.3)$	$f_{12}=D_{12}-D_{12\text{-}NEUTRAL}$	*IRISD*
close_t_r_eyelid (F_{20}) – *close_b_r_eyelid* (F_{22})	$D_{13}=s(3.2,3.4)$	$f_{13}=D_{13}-D_{13\text{-}NEUTRAL}$	*IRISD*
stretch_l_cornerlip (F_6) (*stretch_l_cornerlip_o*)(F_{53}) – *stretch_r_cornerlip* (F_7) (*stretch_r_cornerlip_o*)(F_{54})	$D_{14}=s(8.4,8.3)$	$f_{14}=D_{14}-D_{14\text{-}NEUTRAL}$	*MW*
squeeze_l_eyebrow (F_{37}) AND *squeeze_r_eyebrow* (F_{38})	$D_{15}=s(4.6,4.5)$	$f_{15}=D_{15\text{-}NEUTRAL}-D_{15}$	*ES*

Table 2: *Quantitative FAPs 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 its in neutral position*

It should be noted that not all FAPs included in the vocabularies of Table 1 can be modeled by distances between facial protuberant points (e.g. *raise_b_lip_lm_o*, *lower_t_lip_lm_o*). In such cases the corresponding FAPs are retained in the vocabulary and their ranges of variation are experimentally defined based on facial animations. Moreover, some features serve for the estimation of range of variation of more than one FAP (e.g. features f_{12}-f_{15}).

H. Creation of profiles

We have created several profiles for the archetypal expressions. Every *expression profile* has been created by the selection of a set of FAPs coupled with the appropriate ranges of variation and its animation produces the selected emotion.

In order to define exact profiles for the archetypal expressions, we combine the following steps:

(a) Definition of subsets of candidate FAPs for an archetypal expression, by translating the facial features formations proposed by psychological studies [16], [4], [6] to FAPs,

(b) Fortification of the above definition using variations in real sequences and,

(c) Animation of the produced profiles to verify appropriateness of derived representations.

The initial range of variation for the FAPs has been computed as follows: Let $m_{i,j}$ and $s_{i,j}$ be the mean value and standard deviation of FAP F_j for the archetypal expression i (where i={1→Anger, 2→Sadness, 3→Joy, 4→Disgust, 5→Fear, 6→Surprise}), as estimated in [22]. The initial range of variation $X_{i,j}$ of FAP F_j for the expression i is defined as:

$X_{i,j}=[m_{i,j}-s_{i,j}, m_{i,j}+s_{i,j}]$

for bi-directional, and

$X_{i,j}=[max(0, m_{i,j}-s_{i,j}), m_{i,j}+s_{i,j}]$ or $?_{i,j}=[m_{i,j}-s_{i,j}, min(0, m_{i,j}+s_{i,j})]$

for unidirectional FAPs [14].

For example, the emotion group *fear* also contains *worry* and *terror* [22] which can be synthesized by reducing or increasing the intensities of the employed FAPs, respectively.

Another example is this of emotion *guilty* which is an intermediate expression between *afraid* and *sad*.

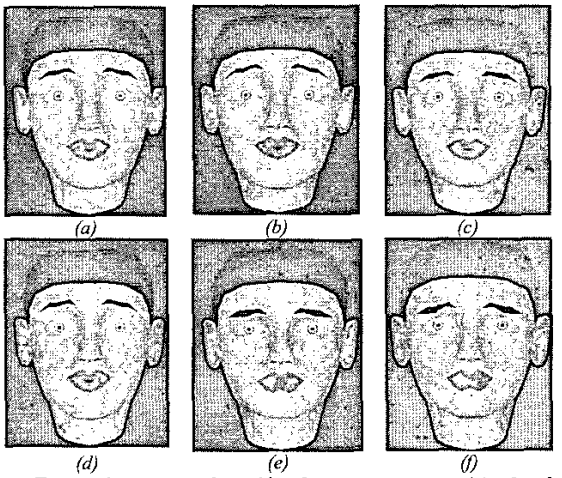

Figure 2: *Animated profiles for emotion terms (a) afraid, (b) terrified, (c) worried, (d) afraid, (e) guilty and (f) sad.*

Figures 2(a)-(c) show the resulting profiles for the terms *terrified* and *worried* emerged by the one of the profiles of *afraid*. Figures 2(d)-(f) show the resulting profiles for the terms *afraid*, *sad* and for the expression lying between them – *guilty*. The profile of *guilty* has been created by the combination of the profiles *afraid* and *sad*. The FAP values that we used are the median ones of the corresponding ranges of variation.

I. Rule based emotion analysis

Let us consider as input to the emotion analysis sub-system a 15-element length feature vector $\boldsymbol{f}$ that corresponds to the 15 features f_i shown in Table 2. Gestures are utilized to support the outcome of this subsystem, since in most cases they are too ambiguous to indicate a positive response. Besides this, quantitative features derived from hand segment tracking are mapped to the emotional space parameters. More specifically, speed and amplitude of motion fortify the position of an observed emotion along the positive activation axis; for example, *satisfaction* turns to *joy* or even to *exhilaration*, as the speed and amplitude of clapping increases.

The particular values of $\boldsymbol{f}$ can be rendered to FAP values as shown in the same table (see also paragraph 3.1) resulting in an input vector $\boldsymbol{G}$. The elements of $\boldsymbol{G}$ express the observed

values of the corresponding involved FAPs; for example G_1 refers to the value of F_{37}.

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 $\mu_{i,j}^{(k)}$. 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 $\boldsymbol{G}$, facial state corresponds to profile $P_i^{(k)}$ and emotion *i* respectively, are computed through the following equations:

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

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 $A_{i,j}^{(k)}$.

If a final decision about the observed emotion has to be made then the following equation is used: $q = \arg\max_i b_i$

It is observed that 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 $u(a,b)=\max(a,b)$.

VI. CONCLUSIONS – FUTURE WORK

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 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. The input to these rules are 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 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. The presented method can be combined with visual elements, e.g. gestures, and extended to a multimodal system, which may be used for the creation of sensitive artificial listener (SAL), e.g. a program which will answer like human being (like ELIZA).

REFERENCES

[1] A. Kendon, How gestures can become like words, in Potyatos, F. (ed), *Crosscultural perspectives in nonverbal communication*, pp. 131-141, Toronto, Canada, Hogrefe, 1988.

[2] A. Wexelblat, An approach to natural gesture in virtual environments, *ACM Transactions on Computer-Human Interaction*, vol. 2, iss. 3, pp. 179 – 200, 1995.

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

[4] F. Parke and K. Waters, *Computer Facial Animation*, A K Peters, 1996

[5] F. Quek, "*Unencumbered gesture interaction*," *IEEE Multimedia*, vol. 3. no. 3, pp. 36-47, 1996.

[6] G. Faigin, "*The Artist's Complete Guide to Facial Expressions*", Watson-Guptill, New York, 1990

[7] J. Bates, The role of emotion in believable agents, *Communications of the ACM*, 37(7):122-125, 1992.

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

[9] J. N. Bassili, '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, 1979.

[10] J.J. Kuch and T.S. Huang, "Vision-based hand modeling and tracking for virtual teleconferencing and telecollaboration," in *Proc. IEEE Int. Conf. Computer Vision*, June 1995, pp. 666-671.

[11] 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 the Electronic Imaging 2000 Conference of SPIE,* San Jose, USA, Jan. 2000, pp 443-450.

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

[13] M. Preda, F. Prêteux, Advanced animation framework for virtual characters within the MPEG-4 standard, in *Proc. of the International Conference on Image Processing*, New York, Sept. 2002, pp.22-25.

[14] A. M. Tekalp, J. Ostermann, "Face and 2-D mesh animation in MPEG-4", *Signal Processing: Image Communication*, Vol. 15, No. 4-5, pp. 387-421, January 2000.

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

[16] P. Ekman, "Facial expression and Emotion," *Am. Psychologist*, vol. 48 pp. 384-392, 1993

[17] 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*, Jan.2001

[18] R. Plutchik, "*Emotion: A psychoevolutionary synthesis*", Harper and Row, New York, USA, 1980.

[19] R. Picard, *Affective Computing*, MIT Press, Cambridge, MA, 1997.

[20] T. Moeslund, E. Granum, "Pose Estimation of a Human Arm using Kinematic Constraints", *in Proc. of the 12th Scandinavian conference in image analysis*, Bergen, Norway, June 2001, pp 1-8.

[21] C. M. Whissel, "The dictionary of affect in language", in R. Plutchnik and H. Kellerman (Eds) *Emotion: Theory, research and experience: vol 4, The measurement of emotions,* Academic Press, New York, 1989, pp. 113-131.

[22] 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.

ConTEL 2003, ISBN: 953-184-052-0