EMOTION SYNTHESIS IN VIRTUAL ENVIRONMENTS

Amaryllis Raouzaiou, Kostas Karpouzis and Stefanos Kollias
Image, Video and multimedia Systems Laboratory, National Technical University of Athens,
9, Heroon Politechniou street, 15773, Zographou, Athens, Greece
Email: {araouz, kkarpou}@image.ntua.gr, stefanos@cs.ntua.gr

Keywords: MPEG-4 facial animation, facial expressions, emotion synthesis

Abstract: Man-Machine Interaction (MMI) systems that utilize multimodal information about users' current emotional state are presently at the forefront of interest of the computer vision and artificial intelligence communities. Interfaces with human faces expressing emotions may 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, emotion synthesis can enhance the atmosphere of a virtual environment 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 to synthesize expressions using a reduced representation, suitable for networked and lightweight applications.

1 INTRODUCTION

Current information processing and visualization systems are capable of offering advanced and intuitive means of receiving input and communicating output to their users. As a result, Man-Machine Interaction (MMI) systems that utilize multimodal information about their users' current emotional state are presently at the forefront of interest of the computer vision and artificial intelligence communities. Such interfaces give the opportunity to less technology-aware individuals, as well as handicapped people, to use computers more efficiently and thus overcome related fears and preconceptions.

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 (Cowie, Douglas-Cowie, Tsapatsoulis, Votsis, Kollias, Fellenz & Taylor, 2001) has shown that even this approach can benefit from taking into account multimodal information.

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. 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.

The real world actions of a human can be transferred into a virtual environment through a representative (avatar), while the virtual world perceives these actions and corresponds through respective system avatars who can express their emotions using human-like expressions and gestures.

In this paper we describe an approach to synthesize expressions via the tools provided in the MPEG-4 standard (Preda & Preteux, 2002) based on real measurements and on universally accepted assumptions of their meaning. These assumptions are based on established psychological studies, as well as empirical analysis of actual video footage from human-computer interaction sessions and human-to-human dialogues. The results of the synthesis process can then be applied to avatars, so as to convey the communicated messages more vividly than plain textual information or simply to make interaction more lifelike.

2 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, especially for the body, for which the animation is much more complex. The FBA part can be also combined with multimodal input (e.g. linguistic and paralinguistic speech analysis).

2.1 Facial Animation

MPEG-4 specifies 84 feature points on the neutral face, which provide spatial reference for FAPs definition. The FAP set contains two high-level parameters, visemes and expressions. In particular, the Facial Definition Parameter (FDP) and the Facial Animation Parameter (FAP) set were designed in the MPEG-4 framework to allow the definition of a facial shape and texture, eliminating the need for specifying the topology of the underlying geometry, through FDPs, and the animation of faces reproducing expressions, emotions and speech pronunciation, through FAPs. 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.

2.2 Body Animation

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. BAPs 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.

BBA (Bone Based Animation)

The MPEG-4 BBA offers a standardized interchange format extending the MPEG-4 FBA (Preda & Preteux, 2002). 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.

3 EMOTION REPRESENTATION

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 (Whissel, 1989) is a representation that is both simple and capable of capturing a wide range of significant issues in emotion. It rests on a simplified treatment of two key themes:

- *Valence*: The clearest common element of emotional states is that the person is materially influenced by feelings that are ‘valenced’, i.e. they are centrally concerned with positive or negative evaluations of people or things or events. The link between emotion and valencing is widely agreed
- *Activation leve*l: Research has 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 acti-

vation 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 (Whissel, 1989).

A surprising amount of emotional discourse can be captured in terms of activation-emotion space. Perceived fullblown emotions are not evenly distributed in activation-emotion 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, identifying the center as a natural origin has several implications. Emotional strength can be measured as the distance from the origin to a given point in activation-evaluation space. The concept of a full-blown emotion can then be translated roughly as a state where emotional strength has passed a certain limit. An interesting implication is that strong emotions are more sharply distinct from each other than weaker emotions with the same emotional orientation. A related extension is to think of primary or basic emotions as cardinal points on the periphery of an emotion circle. Plutchik 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.

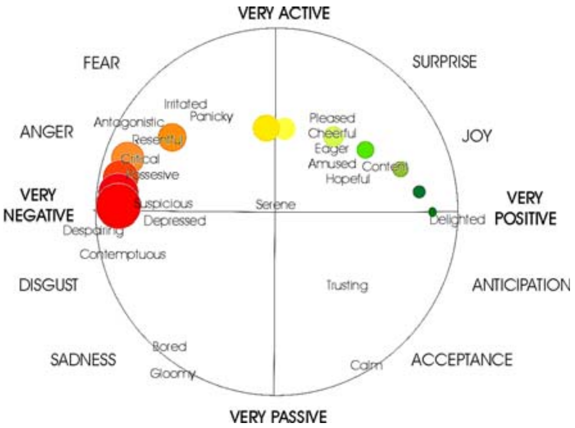

Figure 1: The Activation-emotion space

4 FACIAL EXPRESSIONS

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 twenty 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.

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 (Karpouzis, Tsapatsoulis & Kollias, 2000).

Facial features can be viewed (Cowie et al., 2001) 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 (Bassili, 1979), 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.

In general, facial expressions and emotions are described by a set of measurements and transformations that can be considered atomic with respect to the MPEG-4 standard; in this way, one can describe both the anatomy of a human face –basically through FDPs, as well as animation parameters, with groups of distinct tokens, eliminating the need for specifying the topology of the underlying geometry. These tokens can then be mapped to automatically detected measurements and indications of motion on a video sequence and, thus, help to approximate a real expression conveyed by the subject by means of a synthetic one.

5 GESTURES AND POSTURES

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 one's hands over their 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.

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 (Kuch & Huang, 1995), kinematical structure (Lin, Wu & Huang, 200), dynamics (Quek, 1996), (Wilson & Bobick, 1998) and semantics.

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 (Wexelblat, 1995). 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. This relation is also taken into account in (Kendon, 1988) and provides a positioning of gestures along a continuous space.

6 FROM FEATURES TO SYMBOLS

6.1 Face

In order to estimate the users' emotional state in a MMI context, we must first describe the six archetypal expressions (joy, sadness, anger, fear, disgust, surprise) 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 (Ekman, 1993), (Parke, 1996), (Faigin, 1990), 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 (Karpouzis, Tsapatsoulis & Kollias, 2000). 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 (Tekalp & Ostermann, 2000), 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 (Raouzaiou, Tsapatsoulis, Karpouzis & Kollias, 2002). 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 for the description	Utilized feature
Squeeze_l_eyebrow (F_{37})	$D_1=s(4.5,3.11)$	$f_1=D_{1\text{-}NEUTRAL}-D_1$
Squeeze_r_eyebrow (F_{38})	$D_2=s(4.6,3.8)$	$f_2=D_{2\text{-}NEUTRAL}-D_2$
Lower_t_midlip (F_4)	$D_3=s(9.3,8.1)$	$f_3=D_3-D_{3\text{-}NEUTRAL}$
Raise_b_midlip (F_5)	$D_4=s(9.3,8.2)$	$f_4=D_{4\text{-}NEUTRAL}-D_4$
Raise_l_i_eyebrow (F_{31})	$D_5=s(4.1,3.11)$	$f_5=D_5-D_{5\text{-}NEUTRAL}$
Raise_r_i_eyebrow (F_{32})	$D_6=s(4.2,3.8)$	$f_6=D_6-D_{6\text{-}NEUTRAL}$

Raise_l_o_eyebrow *(F_{35})*	$D_7=s(4.5,3.7)$	$f_7=D_7-D_{7\text{-NEUTRAL}}$
Raise_r_o_eyebrow *(F_{36})*	$D_8=s(4.6,3.12)$	$f_8=D_8-D_{8\text{-NEUTRAL}}$
Raise_l_m_eyebrow *(F_{33})*	$D_9=s(4.3,3.7)$	$f_9=D_9-D_{9\text{-NEUTRAL}}$
Raise_r_m_eyebrow *(F_{34})*	$D_{10}=s(4.4,3.12)$	$f_{10}=D_{10}-D_{10\text{-NEUTRAL}}$
Open_jaw (F_3)	$D_{11}=s(8.1,8.2)$	$f_{11}=D_{11}-D_{11\text{-NEUTRAL}}$
close_t_l_eyelid (F_{19}) – *close_b_l_eyelid (F_{21})*	$D_{12}=s(3.1,3.3)$	$f_{12}=D_{12}-D_{12\text{-NEUTRAL}}$
close_t_r_eyelid (F_{20}) – *close_b_r_eyelid (F_{22})*	$D_{13}=s(3.2,3.4)$	$f_{13}=D_{13}-D_{13\text{-NEUTRAL}}$
stretch_l_cornerlip (F_6) (stretch_l_cornerlip_o)(F_{53}) – *stretch_r_cornerlip (F_7) (stretch_r_cornerlip_o)(F_{54})*	$D_{14}=s(8.4,8.3)$	$f_{14}=D_{14}-D_{14\text{-NEUTRAL}}$
squeeze_l_eyebrow (F_{37}) AND squeeze_r_eyebrow (F_{38})	$D_{15}=s(4.6,4.5)$	$f_{15}=D_{15\text{-NEUTRAL}}-D_{15}$

Table 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

For our experiments on setting the archetypal expression profiles, we used the face model developed by the European Project *ACTS MoMuSys*, being freely available at the website http://www.iso.ch/ittf. Table 3 shows examples of profiles of the archetypal expression fear (Raouzaiou, Tsapatsoulis, Karpouzis & Kollias, 2002).

Figure 2 shows some examples of animated profiles. Fig. 2(a) shows a particular profile for the archetypal expression *anger*, while Fig. 2(b) and (c) show alternative profiles of the same expression. The difference between them is due to FAP intensities. Difference in FAP intensities is also shown in Figures 2(d) and (e), both illustrating the same profile of expression *surprise*. Finally Figure 2(f) shows an example of a profile of the expression *joy*.

Profiles	**FAPs and Range of Variation**
Fear (*$P_F^{(0)}$*)	$F_3 \in [102,480]$,$F_5 \in [83,353]$,$F_{19} \in [118,370]$, $F_{20} \in [121,377]$,$F_{21} \in [118,370]$, $F_{22} \in [121,377]$, $F_{31} \in [35,173]$,$F_{32} \in [39,183]$, $F_{33} \in [14,130]$, $F_{34} \in [15,135]$
$P_F^{(1)}$	$F_3 \in [400,560]$,$F_5 \in [333,373]$,$F_{19} \in [-400,-340]$,$F_{20} \in [-407,-347]$,$F_{21} \in [-400,-340]$,$F_{22} \in [-407,-347]$
$P_F^{(2)}$	$F_3 \in [400,560]$,$F_5 \in [-240,-160]$,$F_{19} \in [-630,-570]$,$F_{20} \in [-630,-570]$,$F_{21} \in [-630,-570]$,$F_{22} \in [-630,-570]$,$F_{31} \in [260,340]$,$F_{32} \in [260,340]$,$F_{33} \in [160,240]$,$F_{34} \in [160,240]$,$F_{35} \in [60,140]$,$F_{36} \in [60,140]$

Table 3: Profiles for the Archetypal Expression Fear

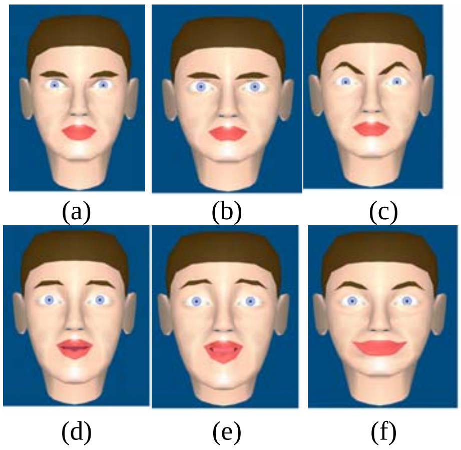

Figure 2: Examples of animated profile: (a)-(c) Anger, (d)-(e) Surprise, (f) Joy

Creating Profiles for Expressions Belonging to the Same Universal Emotion Category

As a general rule, one can define six general categories, each characterized by an archetypal emotion; within each of these categories, intermediate expressions are described by different emotional intensities, as well as minor variation in expression details. From the synthetic point of view, emotions belonging to the same category can be rendered by animating the same FAPs using different intensities. In the case of expression profiles, this affect the range of variation of the corresponding FAPs which is appropriately translated; the fuzziness introduced by the varying scale of FAP intensities provides mildly differentiated output in similar situations. This ensures that the synthesis will not render "robot-like" animation, but drastically more realistic results.

For example, the emotion group *fear* also contains *worry* and *terror* (Raouzaiou et al., 2002), synthesized by reducing or increasing the intensities of the employed FAPs, respectively.

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 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 $\sigma_{i,j}$ 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 (Raouzaiou et al., 2002) . The initial range of variation $X_{i,j}$ of FAP F_j for the expression i is defined as:

$$X_{i,j}=[m_{i,j}-\sigma_{i,j}\,,\, m_{i,j}+\sigma_{i,j}]. \qquad \textbf{(1)}$$

for bi-directional, and

$$X_{i,j}=[max(0,\, m_{i,j}-\sigma_{i,j}),\, m_{i,j}+\sigma_{i,j}] \text{ or} \qquad \textbf{(2)}$$
$$X_{i,j}=[\, m_{i,j}-\sigma_{i,j}\,,\, min(0,\, m_{i,j}+\sigma_{i,j})].$$

for unidirectional FAPs.

For example, the emotion group *fear* also contains *worry* and *terror* (Raouzaiou et al., 2002) which can be synthesized by reducing or increasing the intensities of the employed FAPs, respectively.

Emotion term	**Profile**
Afraid	$F_3 \in [400,560]$, $F_5 \in [-240,-160]$, $F_{19} \in [-630,-570]$, $F_{20} \in [-630,-570]$, $F_{21} \in [-630,-570]$, $F_{22} \in [-630,-570]$,$F_{31} \in [260,340]$, $F_{32} \in [260,340]$, $F_{33} \in [160,240]$, $F_{34} \in [160,240]$, $F_{35} \in [60,140]$, $F_{36} \in [60,140]$
Terrified	$F_3 \in [520,730]$, $F_5 \in [-310,-210]$, $F_{19} \in [-820,-740]$, $F_{20} \in [-820,-740]$, $F_{21} \in [-820,-740]$, $F_{22} \in [-820,-740]$, $F_{31} \in [340,440]$, $F_{32} \in [340,440]$, $F_{33} \in [210,310]$, $F_{34} \in [210,310]$, $F_{35} \in [80,180]$, $F_{36} \in [80,180]$
Worried	$F_3 \in [320,450]$, $F_5 \in [-190,-130]$, $F_{19} \in [-500,-450]$, $F_{20} \in [-500,-450]$, $F_{21} \in [-500,-450]$, $F_{22} \in [-500,-450]$, $F_{31} \in [210,270]$, $F_{32} \in [210,270]$, $F_{33} \in [130,190]$, $F_{34} \in [130,190]$, $F_{35} \in [50,110]$, $F_{36} \in [50,110]$

Table 4: Created profiles for the emotions terror and worry

Table 4 and Figures 3(a)-(c) show the resulting profiles for the terms *terrified* and *worried* emerged by the one of the profiles of *afraid*. The FAP values that we used are the median ones of the corresponding ranges of variation.

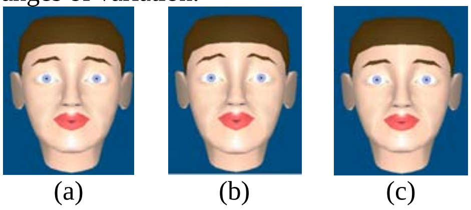
(a) (b) (c)

Figure 3: Animated profiles for (a) afraid, (b) terrified (c) worried

6.2 Gesture Classification

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.

Table 5 shows the correlation between some detectable gestures with the six archetypal expressions.

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

Table 5: Correlation between gestures and emotional states

Given a particular context of interaction, gesture classes corresponding to the same emotional are combined in a "logical OR" form. Table 1 shows that a particular gesture may correspond to more than one gesture classes carrying different affective meaning. For example, if the examined gesture is *clapping*, detection of high frequency indicates *joy*, but a *clapping* of low frequency may express irony and can reinforce a possible detection of the facial expression *disgust*.

Animation of gestures is realized using the 3D model of the software package *Poser*, edition 4 of CuriousLabs Company. This model has separate parts for each moving part of the body. The Poser model interacts with the controls in Poser and has joints that move realistically, as in real person. Poser adds joint parameters to each body part. This allows us to manipulate the figure based on those parameters. We can control the arm, the head, the hand of the model by filling the appropriate parameters; to do this a mapping from BAPs to Poser parameters is necessary. We did this mapping mainly experimentally; the relationship between BAPs and Poser parameters is more or less straightforward.

Figure 4 shows some frames of the animation created using the Poser software package for the gesture "lift of the hand" in the variation which expresses *sadness*.

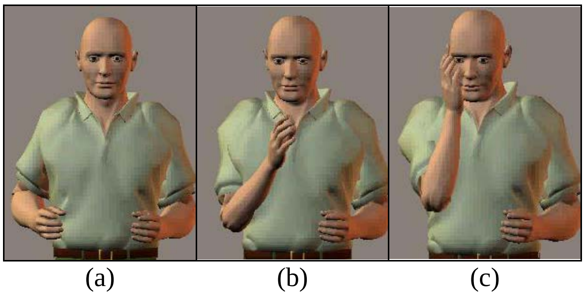

Figure 4: Frames from the animation of the gesture "lift of the hand"

7 CONCLUSIONS

Expression synthesis is a great means of improving HCI applications, since it provides a powerful and universal means of expression and interaction. In this paper we presented a method of synthesizing realistic expressions using lightweight representations. This method employs concepts included in established standards, such as MPEG-4, which are widely supported in modern computers and standalone devices.

REFERENCES

Kendon, A, 1988. How gestures can become like words. In *Crosscultural perspectives in nonverbal communication.* Potyatos, F. (ed.). Hogrefe, Toronto, Canada.

Wexelblat, A., 1995. An approach to natural gesture in virtual environments. In *ACM Transactions on Computer-Human Interaction*, Vol. 2, iss. 3.

Parke, F., Waters, K., 1996. *Computer Facial Animation*. A K Peters.

Quek, F., 1996. Unencumbered gesture interaction. In *IEEE Multimedia*, Vol. 3. no. 3.

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

Lin, J., Wu, Y., Huang, T.S., 2000. Modeling human hand constraints. In *Proc. Workshop on Human Motion*.

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*.

Kuch, J. J., Huang, T. S., 1995. Vision-based hand modeling and tracking for virtual teleconferencing and telecollaboration. In *Proc. IEEE Int. Conf. Computer Vision.*

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

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

Preda, M., Prêteux, F., 2002. Advanced animation framework for virtual characters within the MPEG-4 standard. In *Proc. of the International Conference on Image Processing*. Rochester, NY.

Tekalp, M., Ostermann, J., 2000. Face and 2-D mesh animation in MPEG-4.In *Image Communication Journal*, Vol.15, Nos. 4-5.

Ekman, P., Friesen, W., 1978. The Facial Action Coding System. In *Consulting Psychologists Press*. San Francisco, CA.

Ekman, P., 1993. Facial expression and Emotion. In *Am. Psychologist*, Vol. 48.

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

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

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

Wilson, A., Bobick, A., 1998. Recognition and interpretation of parametric gesture. In *Proc. IEEE Int. Conf. Computer Vision*.

Wu, Y., Huang, T.S., 2001.Hand modeling, analysis, and recognition for vision-based human computer interaction. In *IEEE Signal Processing Magazine.* Vol. 18, iss. 3.

Raouzaiou, A., Tsapatsoulis, N., Karpouzis, K., Kollias, S., 2002. Parameterized facial expression synthesis based on MPEG-4. In *EURASIP Journal on Applied Signal Processing*. Vol. 2002, No. 10. Hindawi Publishing Corporation.