Emotion Synthesis in the MPEG-4 Framework

A. Raouzaiou, K. Karpouzis and S. Kollias
Image, Video and Multimedia Systems Laboratory,
School of Electrical and Computer Engineering
National Technical University of Athens
Athens, Greece
araouz@image.ntua.gr

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. A lifelike human face can enhance interactive applications by providing straightforward feedback to and from the users and stimulating emotional responses from them. 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.**

Keywords—facial animation, emotion synthesis, MPEG-4

I. INTRODUCTION

Research in facial expression analysis and synthesis has mainly concentrated on archetypal emotions. In particular, sadness, anger, joy, fear, disgust and surprise are categories of emotions that attracted most of the interest in human computer interaction environments.

Moreover, the MPEG-4 indicates an alternative way of modeling facial expressions and the underlying emotions, which is strongly influenced from neurophysiological and psychological studies (FAPs). The adoption of token-based animation in the MPEG-4 framework [3] benefits the definition of emotional states, since the extraction of simple, symbolic parameters is more appropriate to analyze, as well as synthesize facial expression and hand gestures.

In this paper we describe an approach to synthesize expressions, including intermediate ones, via the tools provided in the MPEG-4 standard based on real measurements and on universally accepted assumptions of their meaning, taking into account results of Whissel's study [5]. 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.

II. 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.

Activation-emotion space [5] is a representation that is both simple and capable of capturing a wide range of significant issues in emotion. 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 [5].

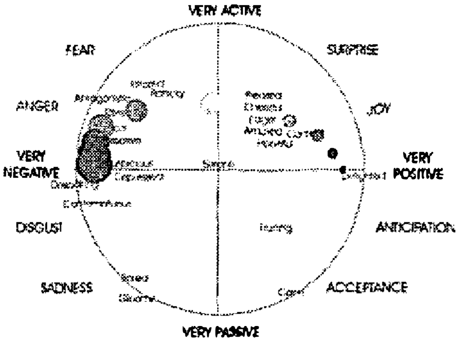


Figure 1. The Activation-emotion space

III. FACE AND BODY ANIMATION IN MPEG-4

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).

A. 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

0-7803-8578-0/04/$20.00 ©2004 IEEE

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.

B. *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 [3]. 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.

IV. Facial Expressions

In order to model an 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 representations [3] 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 [4], [2], we have described the six archetypal expressions using MPEG-4 FAPs [6]; the description for *sadness* is illustrated in Table I. In general, these expressions can be uniformly recognized across cultures and are therefore invaluable in trying to analyze the users' emotional state.

TABLE I. FAPs Vocabulary for Archetypal Expression Sadness

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})*

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 [6]. 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 [7], 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. The units for f_i are identical to those corresponding to FAPs, even in cases where no one-to-one relation exists.

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 II shows examples of profiles of the archetypal expression fear [6].

Fig. 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.

TABLE II. Profiles for the Archetypal Expression Fear

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

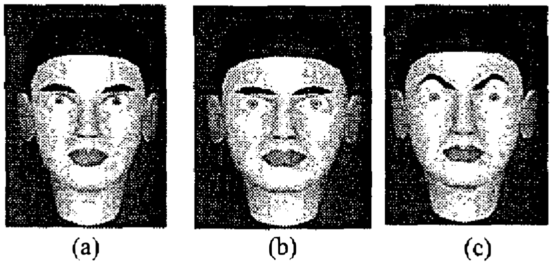


Figure 2. Examples of animated profile: (a)-(c) Anger

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 does not render "robot-like" animation, but drastically more realistic results. For example, the emotion group *fear* also contains *worry* and *terror* [6], 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 [6]. The initial range of variation $X_{i,j}$ of FAP F_j for the expression i is defined as:

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

for bi-directional, and

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

for unidirectional FAPs.

For example, the emotion group *fear* also contains *worry* and *terror* [6] which can be synthesized by reducing or increasing the intensities of the employed FAPs, respectively.

TABLE III. CREATED PROFILES FOR THE EMOTIONS TERROR AND WORRY

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	$F_3 \in [320,450]$, $F_5 \in [-190,-130]$, $F_{19} \in [-500,-450]$, $F_{20} \in [-500,-450]$, $F_{21} \in [-500,-450]$, $F_{22} \in [-500,-450]$,$F_{31} \in [210,270]$, $F_{32} \in [210,270]$,$F_{33} \in [130,190]$, $F_{34} \in [130,190]$, $F_{35} \in [50,110]$, $F_{36} \in [50,110]$

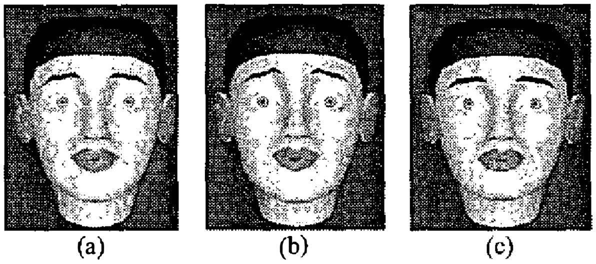


Figure 3. Animated profiles for (a) afraid, (b) terrified (c) worried

Table III and Fig. 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.

V. GESTURES AND POSTURES

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 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.

The low level results of the approach can be extended, taking into account that hand gestures are a powerful expressive means [1]. 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.

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 considerati•n joints' limitations and independent degrees of freedom in the skeleton model of the different body parts.

Gesture Categories

Gestures are utilized to support the synthesis of the facial expression, since in most cases a facial expression is too ambiguous to indicate a particular emotion [8]. However, in a given c•ntext 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.

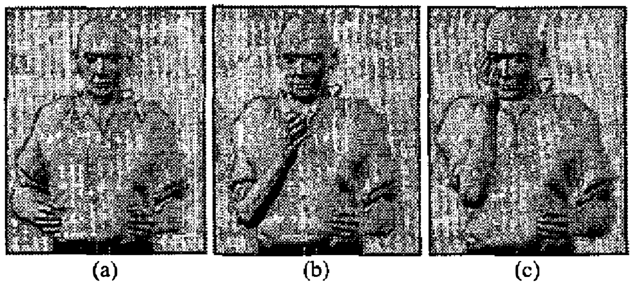
Table IV shows the correlati•n between some gestures with the six archetypal expressions.

TABLE IV. CORRELATION BETWEEN GESTURES AN• EMOTIONAL STATES

Emotion	Gesture Class
Joy	*hand clapping-high frequency*
Sadness	*hands over the head-p•sture*
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*

Animation of gestures is realized using the 3D model of the software package *Poser*, edition 4 of Curi•usLabs 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.

Fig. 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*.



(a) (b) (c)

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

VI. 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 modem computers and standalone devices.

REFERENCES

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

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

[3] M. Preda and F. Prêteux, "Advanced animation framework for virtual characters within the MPEG-4 standard", *Proc. of the Intl. Conference on Image Processing*. Rochester, NY, 2002.

[4] P. Ekman, "Facial expression an• Emotion," *Am. Psychologist*, Vol. 48, pp.384-392, 1993.

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

[6] 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. Hindawi Publishing Corporation, 2002.

[7] M. Tekalp and J. Ostermann, "Face and 2-D mesh animation in MPEG-4," *Image Communicati•n Journal*, V•l.15, Nos. 4-5, 2000.

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