Pers Ubiquit Comput
DOI 10.1007/s00779-007-0164-1

ORIGINAL ARTICLE

MPEG-4 facial expression synthesis

L. Malatesta · A. Raouzaiou · K. Karpouzis · S. Kollias

Received: 11 July 2006 / Accepted: 15 September 2006

Abstract The current work will 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. Additionally, MPEG-4 facial animation parameters are used in order to evaluate theoretical predictions for intermediate expressions of a given emotion episode, based on Scherer's appraisal theory. MPEG-4 FAPs and action units are combined in modeling the effects of appraisal checks on facial expressions and temporal evolution issues of facial expressions are investigated. The results of the synthesizing process can then be applied to Embodied Conversational Agents (ECAs), rendering their interaction with humans, or other ECAs, more affective.

1 Introduction

Affective computing dictates the importance of creating interfaces which are not only solely limited to the synthetic representation of the face and the human body, but which also expresses feelings through facial expressions, gestures and the body pose. The most significant challenge is the compatibility of an Embodied Conversational Agent (ECA) with MPEG-4 standard and its use in various applications. The use of affective avatars can be applied in many sectors—culture, gaming, e-learning, while their compatibility with the MPEG-4 standard makes it possible for avatars to interact with synthetic objects and to be seamlessly integrated in different scenes.

We will 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 [12]. Starting from a symbolic representation of human emotions, based on their expression via facial expressions, we create profiles not only for the six universal expressions (anger, joy, disgust, fear, sadness, surprise), but also for intermediate ones.

In the current work, MPEG-4 facial animation parameters are also used in order to evaluate theoretical predictions for intermediate expressions of a given emotion episode, based on Scherer's appraisal theory. Scherer's appraisal theory investigates the link between the elicitation of an emotion and the response patterning in facial expression [6]. It predicts intermediate expressions based on sequential appraisal checks and postulates a cumulative effect on the final expression.

The current work aims to investigate this cumulative effect through the synthesis of the temporal evolution of

L. Malatesta (✉) · A. Raouzaiou · K. Karpouzis · S. Kollias
Image, Video and Multimedia Systems Laboratory,
National Technical University of Athens,
9, Heroon Politechniou str., Zografou 15780, Greece
e-mail: lori@image.ece.ntua.gr

A. Raouzaiou
e-mail: araouz@image.ece.ntua.gr

K. Karpouzis
e-mail: kkarpou@softlab.ece.ntua.gr

S. Kollias
e-mail: stefanos@cs.ntua.gr

facial expressions during emotion elicitation based on the mapping of Ekman's Action Units [9] to MPEG-4 FAPs.

2 ECA's facial expression and the component process model

The processes of emotion elicitation and emotion expression constitute central issues in rendering an ECA more affective. Emotion theory offers a variety of models each aspiring to capture the emotion expression process. One would expect that the choice of the modeling approach would be irrelevant to the task at hand and would aim to capture global patterns. Contrary to this intuition, relevant research has shown that the choice of the modeling approach is strongly correlated to the task the agent will be asked to carry out. For example the dimensional approach [10] to emotion modeling is more fitting for the case of emotion recognition, i.e., anger detection. It remains a challenge to identify the emotion model for an ECA that will not be dependant of specific action examples.

By studying the requirements for a naturalistic interaction with an ECA, a very central issue in the approach each model adopts is the temporal evolution of an expression and how it is affected by surrounding stimuli. Scherer's component process model provides predictions for intermediate expressions as well as a prediction for the final emotion expression based on appraisal checks preformed on various specifically defined components. In the current work we are interested in evaluating this theoretical model and in investigating ways in which appraisal check results and the accompanying predictions can become a behaviour metric for ECAs in a dynamic environment.

According to cognitive theories of emotion, emotions are closely related to the situation that is being experienced (or, indeed, imagined) by the agent. Specifically, emotions are connected to mental representations that emphasize key elements of a situation and identify them as being either positive or negative. These representations have generally been called appraisals. An appraisal can be thought of as a model, which is selective and valenced, i.e., highlights key elements of a situation and their values for good or ill [1]. Early examples of this approach can be found in [3, 4]. Appraisals are not necessarily conscious, thus the evaluation processes can occur also by an unconscious way as demonstrated by an important corpus of study in cognitive neuroscience, with different methods as subliminal presentations of stimuli or by clinical neuropsychology (e.g., [5]).

Scherer has developed an appraisal model of emotion in which emotions are conceptualized as the outcome of a fixed sequence of checks [6, 7]. According to Scherer's view, emotion serves an important function as ''…an evolved phylogenetically continuous mechanism that allows increasingly flexible adaptation to environmental contingencies by decoupling stimulus and response and thus creating a latency time for response optimization'' [6].

The appraisal is the sequence of Stimulus Evaluation checks (SECs), which represent the smallest set of criteria necessary to account for the differentiation of main groups of emotional states. These checks are not necessarily binary and are subjective (i.e., they depend on both the appraising individual's perception of and inference about the specific characteristics of the event [6].

The individual SECs can be grouped together in terms of what are called Appraisal Objectives, of which there are four: (1) Relevance Detection: comprising Novelty Check, Intrinsic Pleasantness Check, and Goal Relevance Check; (2) Implication Assessment: comprising Causal Attribution Check, Discrepancy from Expectation Check, Goal/Need Conduciveness Check, and Urgency Check; (3) Coping Potential Determination: comprising Control Check, Power Check, and Adjustment Check (can the event be controlled, if so by how much power do I have to exert control, and if not can I adjust?); (4) Normative Significance Evaluation: comprising Internal Standards Check, and External Standards Check. A major assumption of Scherer's SEC Theory is that the sequence of the checks and of the groups is fixed. However, this does not rule out parallel processing as, in theory, all of the SECs are processed simultaneously.

Representations of emotional states using this model of emotion are explained in terms of cognitive appraisals of the antecedent situation, and these appraisals account for the differentiated nature of emotional responses, individual and temporal differences in emotional responses, and for the range of situations that evoke the same response. Appraisals also make appropriate emotional responses likely, and conflict between automatic, unconscious appraisals and more consciously deliberated ones may explain some of the more irrational aspects of emotions [3].

3 MPEG-4 based representation and the facial action coding system

In the framework of MPEG-4 standard [8], parameters have been specified for Face and Body Animation (FBA) by defining specific Face and Body nodes in the scene graph. MPEG-4 specifies 84 feature points on the neutral face, which provide spatial reference for FAPs definition. The FAP set contains two high-level parameters, visemes and expressions. Most of the techniques for facial animation are based on a well-known system for describing ''all visually distinguishable facial movements'' called the Facial Action Coding System (FACS). FACS is an anatomically oriented coding system, based on the definition of

Table 1 FAPs vocabulary for archetypal expression anger

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})

''Action Units'' (AU) of a face that cause facial movements and tries to distinguish the visually distinguishable facial movements using the knowledge of facial anatomy. An AU could combine the movement of two muscles or work in the reverse way, i.e., split into several muscle movements. MPEG-4 FAPs are strongly related to the AU [11]. Description of archetypal expressions by means of muscle movements and AUs has been the starting point for setting the archetypal expression description through FAPs.

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.

4 Modeling universal 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 make good candidates for describing quantitative facial 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 and from statistical analysis, we have described the six archetypal expressions using MPEG-4 FAPs [10]; the description for *anger* 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.

Table 2 shows examples of profiles of the same archetypal expression.

5 Modeling intermediate expressions

The limited number of studies, carried out by computer scientists and engineers, dealing with emotions other than the archetypal ones, lead us to search in other subject/discipline bibliographies. Psychologists examined a broader set of emotions, but very few of the corresponding studies provide exploitable results to computer graphics and machine vision fields, e.g., Whissel's wheel [12]. The synthesis of intermediate expressions is based on the profiles of the six archetypal expressions [10].

Table 2 Profiles for the archetypal expression anger

Profiles	FAPs and range of variation
Anger ($P_A^{(0)}$)	$F_4 \in [22, 124]$, $F_{31} \in [-131, -25]$, $F_{32} \in [-136, -34]$, $F_{33} \in [-189, -109]$, $F_{34} \in [-183, -105]$, $F_{35} \in [-101, -31]$, $F_{36} \in [-108, -32]$, $F_{37} \in [29,85]$, $F_{38} \in [27,89]$
$P_A^{(1)}$	$F_{19} \in [-330, -200]$, $F_{20} \in [-335, -205]$, $F_{21} \in [200,330]$, $F_{22} \in [205,335]$, $F_{31} \in [-200, -80]$, $F_{32} \in [-194, -74]$, $F_{33} \in [-190, -70]$, $F_{34} = \in [-190, -70]$
$P_A^{(2)}$	F_{19} $[-330, -200]$, $F_{20} \in [-335, -205]$, $F_{21} \in [200,330]$, $F_{22} \in [205,335]$, $F_{31} \in [-200, -80]$, $F_{32} \in [-194, -74]$, $F_{33} \in [70,190]$, $F_{34} \in [70,190]$
$P_A^{(3)}$	$F_{16} \in [45,155]$, $F_{18} \in [45,155]$, $F_{19} \in [-330, -200]$, $F_{20} \in [-330, -200]$, $F_{31} \in [-200, -80]$, $F_{32} \in [-194, -74]$, $F_{33} \in [-190, -70]$, $F_{34} \in [-190, -70]$, $F_{37} \in [65,135]$, $F_{38} \in [65,135]$
$P_A^{(4)}$	$F_{16} \in [-355, -245]$, $F_{18} \in [145,255]$, $F_{19} \in [-330, -200]$, $F_{20} \in [-330, -200]$, $F_{31} \in [-200, -80]$, $F_{32} \in [-194, -74]$, $F_{33} \in [-190, -70]$, $F_{34} \in [-190, -70]$, $F_{37} \in [65,135]$, $F_{38} \in [65,135]$

5.1 Creating profiles for expressions belonging to the same universal emotion category

As a general rule, one can define six general categories, each one characterized by an archetypal emotion. From the synthetic point of view, emotions that belong to the same category can be rendered by animating the same FAPs using different intensities. For example, the emotion group fear also contains worry and terror; these two emotions can be synthesized by reducing or increasing the intensities of the employed FAPs, respectively. In the case of expression profiles, this affects the range of variation of the corresponding FAPs, which is appropriately translated.

Table 3 shows the resulting profiles for the terms *terrified* and *worried* generated by the one of the profiles of *afraid*. The FAP values that we used are the median ones of the corresponding ranges of variation.

5.2 Intermediate expressions lying between universal ones

Creating profiles for emotions that do not clearly belong to a universal category is not straightforward. Apart from estimating the range of variations for FAPs, one should first define the FAPs, which are involved in the particular emotion.

One is able to synthesize intermediate emotions by combining the FAPs employed for the representation of universal ones. In our approach, FAPs that are common in both emotions are retained during synthesis, while emotions used in only one emotion are averaged with the respective neutral position. In the case of mutually exclusive FAPs, averaging of intensities usually favors the most exaggerated of the emotions that are combined, whereas FAPs with contradicting intensities are cancelled out. The rules used to merge the profiles of the archetypal expressions in order to derive intermediate ones can be found at [10].

5.3 Synthesis of universal and intermediate facial expressions

Figure 1 shows some examples of animated profiles. Figure 1a shows a particular profile for the archetypal expression *anger*, while Fig. 1b and c shows alternative profiles of the same expression. The difference between them is due to FAP intensities. The procedure of profiles extraction is described in [10].

Figure 2a–c show the resulting profiles for the terms *terrified* and *worried* generated by the one of the profiles of *afraid*.

As far as intermediate expressions belonging to a different group are concerned, it should be noted that the profiles derived by the described method, have to be animated for testing and correction purposes; the final profiles are those that are approved by experts, e.g., they present an acceptable visual similarity with the requested real emotion.

Using the rules of [10], *depression* (Fig. 3b, c) and *guilt* (Fig. 4b) is animated using *fear* (Fig. 3a, 4a) and *sadness* (Fig. 3c, 4c), *suspicious* (Fig. 5b) using *anger* (Fig. 5a) and *disgust* (Fig. 5c). From Fig. 3b and c, b is approved and c is rejected.

Table 3 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	$F_3 \in [320,450]$, $F_5 \in [-190, -130]$, $F_{19} \in [-500, -450]$, $F_{20} \in [-500,-450]$, $F_{21} \in [-500,-450]$, $F_{22} \in [-500,-450]$, $F_{31} \in [210,270]$, $F_{32} \in [210,270]$, $F_{33} \in [130,190]$, $F_{34} \in [130,190]$, $F_{35} \in [50,110]$, $F_{36} \in [50,110]$

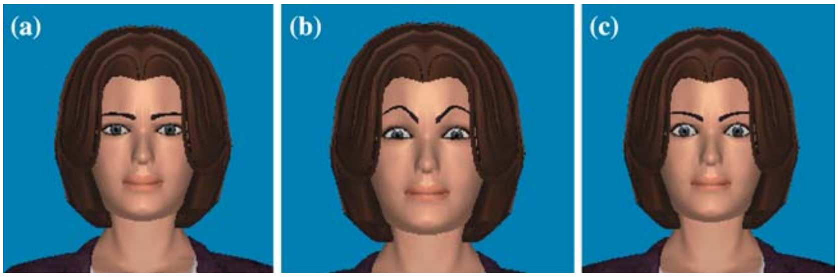

Fig. 1 Examples of animated profile: **a**–**c** anger

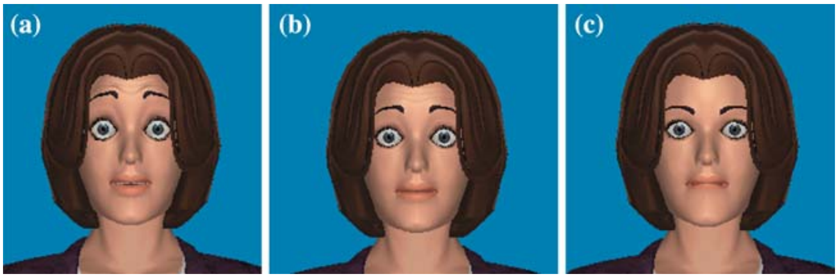

Fig. 2 Animated profiles for **a** terrified, **b** afraid **c** worried

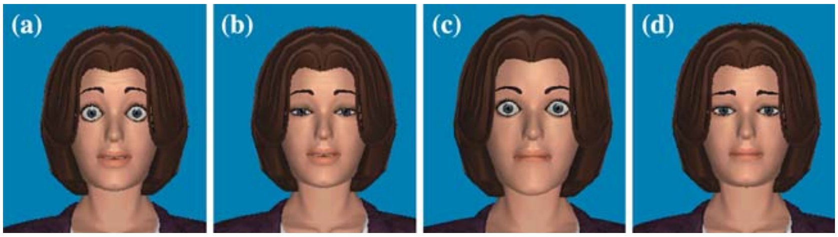

Fig. 3 Profiles for **a** fear, **b**–**c** depressed, **d** sadness

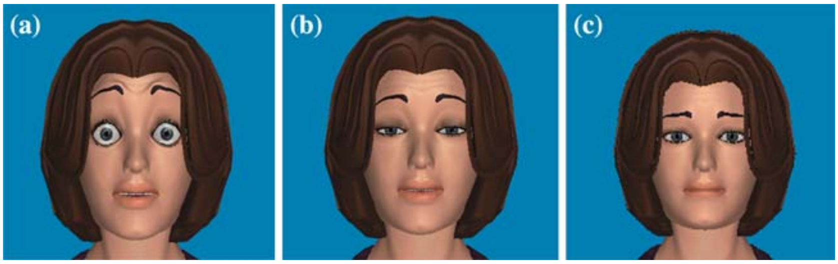

Fig. 4 Profiles for **a** fear, **b** guilt, **c** sadness

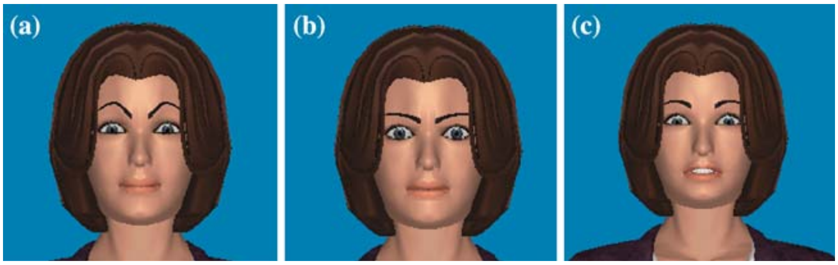

Fig. 5 Profiles for **a** anger, **b** suspicious, **c** disgust

6 Facial expression synthesis based on appraisal theory predictions

Based on the predictions of Scherer's appraisal theory for the intermediate expressions of hot anger and fear, videos animating the transition between the predicted expressions were generated using the GretaPlayer MPEG-4 decoder. The process was based on the mapping of Ekman's Action Units to MPEG-4 FAPs [11]. This approach aims to be the beginning of an attempt to model the effects of appraisal checks on facial expressions, taking advantage of the flexibility and the expressivity the GretaPlayer engine has to offer.

Until recently, most of our work had to do with static images of the apex of an expression, since no videos with satisfactory resolution that would allow the tracking of the evolution of a FAP in successive frames were available. In contrast to a display of static images of the intermediate predictions, in the process of video synthesis, the temporal evolution of the expressions poses various issues on the synthesis procedure. Lacking the information about the track each facial animation parameter follows through time; various methods of transition between the intermediate expressions were investigated.

The appraisal theory predicts a *cumulative effect* of intermediate predictions on the final expression of an emotion. This effect needs empirical investigation in order to determine the appropriate method of animation of the effect. We have identified two major ways of treating the evolution of an expression between the intermediate

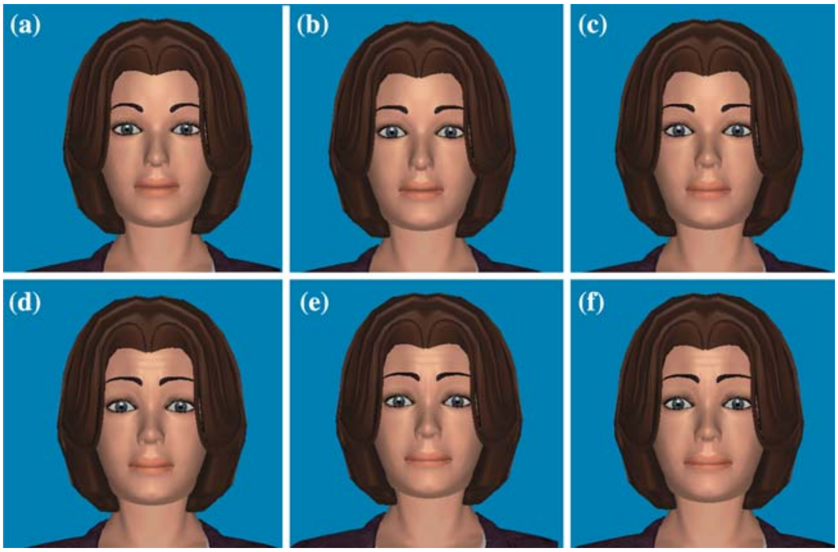

Fig. 6 Intermediate predictions of facial expressions according to Scherer's appraisal theory for the case of fear: **a** neutral, **b** novelty-sudden, **c** unpleasant, **d** discrepant, **e** goal obstractive, **f** low control-final expression-fear. Each expression is derived from the ''addition'' of the previous expression's action units and those of the current one

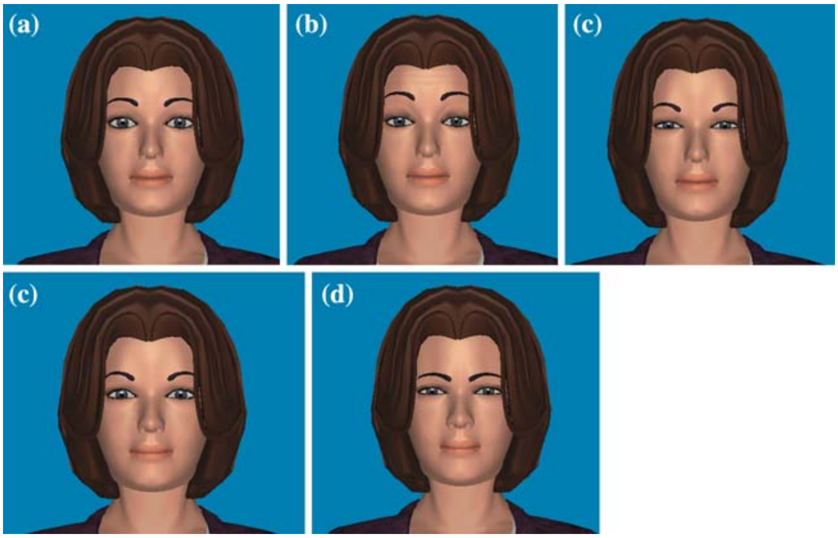

Fig. 7 Intermediate predictions of facial expressions according to Scherer's appraisal theory for the case of hot anger: **a** neutral, **b** novelty-high, **c** goal obstructive, **d** control high/ power high, **e** final expression-hot anger

expression predictions provided by the appraisal checks, an additive animation and a sequential one. They are methods based on principles of computer graphics that require further empirical testing on the naturalness of their outcome. In this preliminary research both approaches were tested in depth, the sequential presentation of intermediate expressions was used in the case of hot anger and the additive presentation of the intermediate expressions was used in the case of fear. Results on a frame level can be seen in Figs. 6 and 7.

In the case of additive animation—as seen in the fear example, each intermediate expression is derived by the addition of the AUs of the current expression to the AUs of the previous appraisal check AUs.

This approach was found to be problematic in the cases when subsequent expressions are constituted of conflicting animations. For example in the case of hot anger the ''novelty high'' intermediate expression, according to the appraisal theory predictions ([2]) includes raised eyebrows among others. The next intermediate prediction is ''goal obstructive'' and predicts lowered eyebrows. This conflict renders the animation problematic and the outcome of a sequential representation is confusing.

In the case of sequential animation—as adopted in the hot anger example, all intermediate expressions are animated in sequence. This could be realized either by interposing the neutral expression between the predictions or by ''tweening'' from one expression to the other keeping the common deformations as the common denominator. The approach containing the neutral expressions between predicted expressions renders the outcome counterintuitive. Overall the tweening approach is friendlier to the eye but is still not perceived as a realistic expression generation. Such conclusions demand further investigation in order to empirically prove such hypotheses. Both expert and simple user evaluation is needed.

7 Conclusion: future work

The synthesis of emotional facial expressions should be used to systematically address the questions of the underlying mechanisms of the emotion elicitation process, as well as the temporal unfolding emotional expression. The results presented in this paper aim to constitute the basis of future research and interdisciplinary collaboration between relevant research groups. Expert opinions as well as specific hypothesis testing are required to back or falsify the current preliminary conclusions. Future work will be comprised of in depth investigation of the temporal evolution issues that arose. More emotional expressions need to be synthesized in order to obtain substantial empirical evidence on the veracity of the appraisal theory predictions in expression synthesis.

Acknowledgments This research is partly supported by the EC Project HUMAINE (IST-507422).

References

1. Picard RW (1997) Affective computing. MIT Press, Cambridge
2. Wehrle T, Kaiser S, Schmidt S, Scherer KR (2000) Studying the dynamics of emotional expression using synthesized facial muscle movements. J Pers Soc Psychol 78(1):105–119
3. Roseman IJ, Smith CA (2001) Appraisal theory: overview, assumptions, varieties, controversies. In: Scherer KR, Schorr A, Johnstone T (eds) Appraisal processes in emotion: theory methods, research. Oxford University Press, Oxford, pp 3–19
4. Ortony A, Clore GL, Collins A (1988) The cognitive structure of emotions. Cambridge University Press, Cambridge
5. Pegna AJ, Khateb A, Lazeyras F, Seghier ML (2004) Discriminating emotional faces without primary visual cortices involves the right amygdala. Nat Neurosci 8(1):24–25
6. Scherer KR (2001) Appraisal considered as a process of multilevel sequential checking. In: Scherer KR, Schorr A, Johnstone T (eds) Appraisal processes in emotion: theory methods, research. Oxford University Press, Oxford, pp 92–129
7. Scherer KR (1984) On the nature and function of emotion: a component process approach. In: Scherer KR, Ekman P (eds) Approaches to emotion. Lawrence Erlbaum Associates, Hillsdale, pp 293–318
8. Tekalp M, Ostermann J (2000) Face and 2-D mesh animation in MPEG-4. Image Commun J 15(4–5):387–421
9. Ekman P (1993) Facial expression and emotion. Am Psychol 48:384–392
10. Raouzaiou A, Tsapatsoulis N, Karpouzis K, Kollias S (2002) Parameterized facial expression synthesis based on MPEG-4. EURASIP J Appl Signal Process 2002(10):1021–1038
11. Raouzaiou A, Caridakis G, Malatesta L, Karpouzis K, Grandjean D, Burkhardt F, Kollias S (2007) Emotion theory and multimodal synthesis of affective ECAs. In: Pelachaud C, Cañamero L (eds) Submitted for publication to Achieving Human-Like Qualities in Interactive Virtual and Physical Humanoids. Special Issue of the International Journal of Humanoid Robotics
12. Whissel CM (1989) The dictionary of affect in language. In: Plutchnik R, Kellerman H (eds) Emotion: theory, research and experience, vol 4. The measurement of emotions. Academic, New York