Dimensional Emotion Driven Facial Expression Synthesis Based on the Multi-Stream DBN Model

Hao Wu[*+], Dongmei Jiang[*+], Yong Zhao[*+],and Hichem Sahli[†⋆]
VUB-NPU Joint Research Group on AVSP
[*]Northwestern Polytechnical University, Xi'an, China
[+]Shaanxi Provincial Key Lab on Speech and Image Information Processing, China
E-mail: jiangdm@nwpu.edu.cn Tel: +86-29-88431532
[†]Vrije Universiteit Brussel, Brussels, Belgium
[⋆] Interuniversity Microelectronics Centre – IMEC, Brussels, Belgium
E-mail: hsahli@vub.ac.be Tel:+32-2-6292916

Abstract—This paper proposes a dynamic Bayesian network (DBN) based MPEG-4 compliant 3D facial animation synthesis method driven by the (Evaluation, Activation) values in the continuous emotion space. For each emotion, a state synchronous DBN model (SS_DBN) is firstly trained using the Cohn-Kanade (CK) database with two streams of inputs: (i) the annotated (Evaluation, Activation) values, and (ii) the extracted Facial Action Parameters (FAPs) of the face image sequences. Then given an input (Evaluation, Activation) sequence, the optimal FAP sequence is estimated via the maximum likelihood estimation (MLE) criterion, and then used to construct the MPEG-4 compliant 3D facial animation. Compared with the state-of-the-art approaches where the mapping between the emotional space and the FAPs has been made empirically, in our approach the mapping is learned and optimized using DBN to fit the input (Evaluation, Activation) sequence. Emotion recognition results on the constructed facial animations, as well as subjective evaluations, show that the proposed method obtains natural facial animations representing well the dynamic process of the emotions from neutral to exaggerate.

I. INTRODUCTION

Embodied Conversational Agents (ECA) requires creating interfaces which are not only limited to the synthetic representation of the face and human body, but also express feelings through facial expressions, gestures and body poses[1]. Affective avatars can be applied in many areas, such as gaming, HCI, e-learning, virtual reality, etc…. The current work devotes to the synthesis of the temporal evolution of facial expressions based on the mapping of (a pair of) Activation-Evaluation[2] to MPEG-4 Facial Action Parameters (FAPs).

In recent years, both image-based rendering and 3D model based facial animations have been proposed for realistic expressive facial animation synthesis. Wu et al.[3] developed a real-time audio-visual Chinese speech synthesizer with a 3D expressive avatar, in which the FAPs for six expressions are manually set by the XfaceEd toolkit[4] to match the facial expressions of the JAFFE database[5]. Other works have addressed the modeling of facial expressions by symbolic (rule-based) approaches. Cassell et al.[6] presented a rule-based automatic system that generates expressions and speech for multiple conversation agents. The system is based on the Facial Action Coding System (FACS) of Ekman[7]. Even though the resulting facial expressions look promising, the system generates always the same expressions in any context. This drawback has been partially overcome in [8] by enlarging and enriching the rules with more details, and modeling variable emotion intensities. A similar modeling is proposed by Bui et al. [9] using a fuzzy rule-based system to map representations of the emotional state of an animated agent onto muscle contraction values for specific facial expressions.

Most of the rule-based systems suffer the drawback of a static generation of facial expressions, due to the fact that the set of rules and their combinations is limited. To overcome these limitations, recently stochastic systems have been proposed. For example, in [10], Hidden Markov Models (HMMs), being able to model time series with uncertainty, have been used for the synthesis of emotional facial expressions during speech. The HMMs were trained on a set of emotion examples with the tracked 3D reflective markers of various facial expressions. More realistic looking facial expressions can be reproduced by modeling the dynamics of human expressions. As an extension to HMM, Zhang et al.[11] proposed to use Dynamic Bayesian Networks (DBNs) for modeling both spatial and dynamic relationships among facial expressions for the analysis and synthesis of the six basic facial expressions. In their approach, they integrate the Action Units (AU)[7] and the FAPs into a DBN to generate the probability distribution of the six facial expressions. For the synthesis, they use the probability distribution of the six facial expressions produced by the analysis, and reconstruct the FAPs and their intensity through a static Bayesian network (BN) to provide quantitative information about the facial expressions and their temporal evolution.

The above cited facial expression synthesis research focused on the basic emotion categories. Other studies have used the Pleasure, Arousal, and Dominance (PAD) 3D-

emotional space[12] to develop rule-based or parametric-based facial expression synthesis systems, in which the mapping between the emotional space and the facial animation parameters has been made empirically. In these approaches, to synthesize facial expressions, emotions are not limited to isolated categories but can be described and quantified along three[13-15] or two[16,17] independent dimensions.

In this research, for each emotion, we model the relationship between the Activation-Evaluation emotional space and the FAPs based on a two stream state synchronous DBN model (SS_DBN). The image sequences of the Cohn-Kanade (CK) facial database[18] are firstly annotated using the Feeltrace toolkit[19] to get the (Evaluation, Activation) labeling of each image, and FAP parameters are extracted based on the detected and tracked facial feature points, then the (Evaluation, Activation) sequences and FAP parameters are input to train the SS_DBN models.

Once the parameters of the SS_DBN models are trained, given an (Evaluation, Activation) curve in the 2D emotional space, the optimal FAP parameters are estimated based on the maximum likelihood estimation (MLE) criterion, which are then used to synthesize the MPEG-4 compliant facial animations. In our experiments, both the emotion recognition experiments and subjective evaluations are done on the synthesized 3D facial animations.

The remainder of the article is organized as follows: In Section II, we describe the labeling of the face images in the 2D emotional space. Section III addresses the extraction of the FAP parameters. Section IV introduces the structure of the SS_DBN model and defines the conditional probability distributions of the nodes. Section V induces the optimal FAP feature learning algorithm based on the maximum likelihood criterion. Section VI discusses our experimental results. Finally, Section VII draws a conclusion for this paper.

II. LABELING IN 2D EMOTIONAL SPACE

In this paper the Cohn-Kanade (CK) facial database[18], with 6 posed emotions (happy, sad, angry, fear, surprise and disgust), has been annotated using the Feeltrace toolkit[19]. Feeltrace allows coders to track the emotional content of a stimulus, as they perceive over time, and move their cursor within the 2-dimensional (Evaluation, Activation) emotion space to rate their impression about the emotional state of the subject. Some examples of the continuous annotations are illustrated in Fig.1.

Since the expressions of the facial image sequences in the database are always from neutral to exaggerate, it is relatively less hard to guarantee the consistency of the labeling of the different sequences with the same emotion.

III. MPEG-4 BASED REPRESENTATION AND FAP GENERATION

The MPEG4 standard describes how to define and animate a human face in a 3D scene[20]. MPEG-4 specifies 84 Feature Points (FPs) on the neutral face to describe the shape of the face model which covers eyes, eyebrows, nose, mouth, tongue, teeth etc. al.

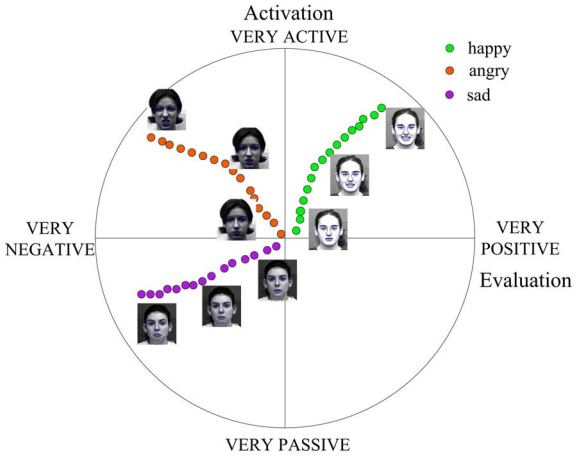

Fig.1 Labeling of facial image sequences in the emotional space

TABLE I
USED FAPS IN THIS PAPER

FAP	Name	Description
3	Open_jaw	Vertical jaw displacement
19	Open_t_l_eyelid	Vertical top left eyelid displacement
21	Close_b_l_eyelid	Vertical bottom left eyelid displacement
31	Raise_l_i_eyebrow	Vertical displacement of right corner of left eyebrow
33	Raise_l_m_eyebrow	Vertical displacement of midpoint between left corner and middle of eyebrow
35	Raise_l_o_eyebrow	Vertical left outer eyebrow displacement
37	Squeeze_l_eyebrow	Horizontal displacement of left eyebrow
51	Lower_t_middle_o	Vertical top middle outer lip displacement
52	Raise_b_midlip_o	Vertical bottom middle outer lip displacement
53	Stretch_r_corner_o	Horizontal displacement of left outer lip corner
59	Raise_l_cornerlip_o	Vertical displacement of left outer lip corner
61	Stretch_l_nose	Horizontal displacement of left corner of nose

The MPEG-4 FPs provide spatial reference for defining (i) 68 Facial Description Parameters (FDP) allowing the definition of a facial shape and texture, as well as eliminating the need for specifying the topology of the underlying geometry, (ii) the Facial Action Parameters (FAP) describing the local movement of the face and hence allowing the animation of faces reproducing expressions, emotions and speech pronunciation.

The FAP set contains two high-level parameters, visemes and expressions, and 66 Low-level FAPs associated with movements of key face zones. All low level FAPs are expressed in terms of facial animation parameter units (FAPU). FAPUs correspond to fractions of distances between some key feature points in a neutral face, e.g., mouth-nose separation, eye separation, etc. MPEG-4 FAPs are strongly

related to the Action Units (AU) and Facial Action Coding System (FACS) describing archetypal expressions by means of muscle movements[7].

For the FAP generation, we use the Constrained Bayesian Tangent Shape Model (CSM)[21] for the detection and tracking, over a facial image sequence, of a shape model defined by 83 facial feature points. In our current implementation, we generate 12 low level FAP parameters (see Table I) out of the 66 low level parameters, extracted directly from the tracked feature points. The considered 12 FAPs are active in facial expressions to characterize the six archetypal facial expressions.

IV. THE STATE SYNCHRONOUS DBN MODEL

The DBN model enables to correlate and associate the continual arriving evidences through temporal dependencies to perform reasoning over time. In our work, for each individual emotion, a two stream state synchronous DBN model (SS_DBN), as shown in Fig.2, is built. It consists of a Prologue part (initialization), a Chunk part that is repeated every time frame, and a closure with an Epilogue part. Every horizontal row of the nodes depicts a separate temporal layer of random variables. The straight arcs represent deterministic conditional probabilities between nodes and the dotted arcs denote random conditional probabilities. The two input streams are the frame-based (Evaluation, Activation) annotations, and the extracted 12 low level FAP parameters of the face images. At each time slice (frame) of the SS_DBN model, the emotion dimensions (Evaluation, Activation) (o^e) and the FAPs (o^v) share the same state variable, i.e. they are forced to be synchronous at the state level.

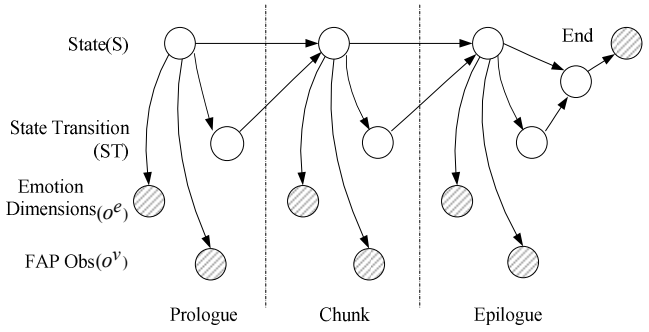

Fig. 2 The state synchronous DBN emotion model (SS_DBN)

The detailed definitions of the nodes are as follows:

- State (S): hidden state of the emotion.
- State Transition (ST): random variable indicating when the current state ends and transits to the next state. $ST = \mathbf{1}$ means that the state is allowed to transit and $ST = \mathbf{0}$ the transition is not allowed.
- Emotion Dimensions (o^e): the (Evaluation, Activation) coordinates of the facial images in the 2D emotional space.
- FAP Obs(o^v): visual observation features, i.e. the 12 key FAPs as defined in Table I.

The conditional probability distributions (CPD) of the key nodes are defined as follows:

1) The CPD of ST is defined as a random probability.

$$P(ST_t = j \mid S_t = i) = \begin{cases} \alpha_{ii} & if \quad j = 0 \\ 1 - \alpha_{ii} & if \quad j = 1 \end{cases} \tag{1}$$

where α_{ii} is a random variable.

2) Suppose the maximum state number is *SM*, the CPD of the state S is defined as

$$p(S_t = i \mid S_{t-1} = j, ST_{t-1} = k) = \begin{cases} 1 & i = j \quad and \quad k = 0 \\ 1 & i = j \quad and \quad j = SM \\ 1 & i = j+1 \ and \quad k = 1 \quad and \quad j < SM \\ 0 & otherwise \end{cases} \tag{2}$$

which means that when the state does not reach the final state and is allowed to transit, the state will change. Otherwise the state will stay on the current state.

3) At each time slice, the probability of the hidden state S emitting the observation feature vectors is defined as a production of Gaussian mixture models (GMMs).

$$p(o_t^e, o_t^v \mid S_t = j) = \prod_{d \in \{e,v\}} \left[\sum_{m=1}^{M} c_{jm}^d N(o_t^d, \mu_{jm}^d, \Theta_{jm}^d) \right]^{w_d} \tag{3}$$

For each input feature stream d ($d \in (e, v)$), c_{jm}^d, μ_{jm}^d and Θ_{jm}^d are the GMM parameters (weight, mean and covariance matrix) of the Gaussian mixture m of the state j, respectively. M is the number of Gaussian mixtures. w_d is the weight adjusting the influence of the stream d, with the constraint $w_e + w_v = 2$. In the training process of the SS_DBN models, w_e and w_v are set to 1, respectively.

For each of the six emotions (happy, sad, angry, fear, surprise and disgust), a SS_DBN model is trained using the labeled (Evaluation, Activation) values and the generated FAPs from the tracked facial feature points. For each feature stream d, a GMM set $\lambda^d = \{c_{jm}^d, \mu_{jm}^d, \Theta_{jm}^d\}$ $(j = 1...SM, m = 1...M)$ is estimated using the Expectation Maximization (EM) algorithm. In our experiments, the number of Gaussian mixtures *M*, as well as the maximum state number *SM*, are set as 4. The training processes of the SS_DBN models are implemented using the Graphical Models Toolkit (GMTK) [22].

V. FACIAL EXPRESSION SYNTHESIS BASED ON THE SS_DBN MODELS

For the synthesis, given an input sequence of (Evaluation, Activation) $o^e = \{o_1^e, ..., o_t^e, ..., o_T^e\}$, we would like to generate the optimal FAP sequence $o^{v'} = \{o_1^{v'}, ..., o_t^{v'}, ..., o_T^{v'}\}$, and then

use these FAP parameters to synthesize a facial animation. This is done based on the Maximum Likelihood (ML) criterion, as follows.

Let ψ_t be the set of all hidden variables at frame t. The probability of an expressional facial image sequence (o^e, o^v) evolving along a hidden variable path $\Psi = (\psi_1, \psi_2, ... \psi_T)$ can be defined as:

$$P\left(O^e, O^v, \Psi \mid \lambda\right) = \prod_{t=1}^{T} p(o_t^e \mid S_t)\, p(o_t^v \mid S_t)\, p\left(\psi_t \mid \psi_{t-1}\right) \quad (4)$$

Given an input (Evaluation, Activation) sequence O^e and the trained model set $\lambda = (\lambda^e, \lambda^v)$, the Maximum Likelihood (ML) criterion is used to find the optimal FAP sequence by iteratively maximizing an auxiliary function $\Omega(\lambda; O^e, O^v, O^{v'})$ defined as:

$$\Omega(\lambda; O^e, O^v, O^{v'}) = \sum_{\Psi \in \Phi} P\left(O^e, O^v, \Psi \mid \lambda\right) \cdot \log\left[P\left(O^e, O^{v'}, \Psi \mid \lambda\right)\right] \quad (5)$$

where $o^{v'}$ is the newly estimated FAP sequence, and o^v is the obtained visual feature sequence in the last iteration, respectively. The optimal FAP $o_t^{v'}$ can be obtained by setting the derivative of $\Omega(\lambda; O^e, O^v, O^{v'})$ with respect to $o_t^{v'}$ equal to zero, $o_t^{v'}$ is then estimated as:

$$o_t^{v'} = \frac{\sum_{\psi_t} P\left(O^e, O^v, \psi_t \mid \lambda\right) \cdot \sum_m c_{\psi_t m}^v \left(\Theta_{\psi_t m}^v\right)^{-1} \mu_{\psi_t m}^v}{\sum_{\psi_t} P\left(O^e, O^v, \psi_t \mid \lambda\right) \cdot \sum_m c_{\psi_t m}^v \left(\Theta_{\psi_t m}^v\right)^{-1}} \quad (6)$$

where $P\left(O^e, O^v, \psi_t \mid \lambda\right)$ is the probability of the facial image sequence (o^e, o^v) passing through ψ_t. We estimate $o_t^{v'}$ by replacing the sum over all possible states of the hidden variables ψ_t, by the sum over their states in the N-Best paths.

Having obtained the 12 key FAP parameters according to the input (Evaluation, Activation) sequence, the other related FAPs, for synthesizing a MPEG-4 facial animation, could be estimated by linear mapping[23]. Finally, we use the Xface open source toolkit[24] to render these parameters.

VI. EXPERIMENTS AND RESULTS

In our experiments, for each of the six emotions, 30 facial image sequences from the Cohn-Kanade database, with their labeled (Evaluation, Activation) (see section II) and generated 12 FAPs (Section III), are used to train the SS_DBN model (Section IV). As testing sequences, we used for each emotion, other 5 facial image sequences which we also annotated within the 2-dimensional (Evaluation and Activation) emotion space using the Feeltrace annotation toolkit[19]. For these sequences, we estimate the optimal FAP parameters based on the trained SS_DBN models, as explained in Section V. This allowed constructing MPEG-4 compliant 3D facial animations.

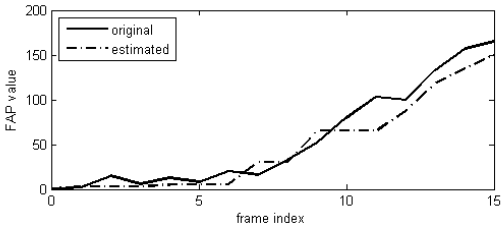

Fig. 3 Temporal dynamics of the FAP parameter: estimated v.s. original

Fig. 3 shows the dynamics of the estimated, as well as the original generated FAP parameters for FAP3 using the procedure of Section III, from a testing facial image sequence, One can notice that the estimated FAP fits well the dynamics of the original FAP.

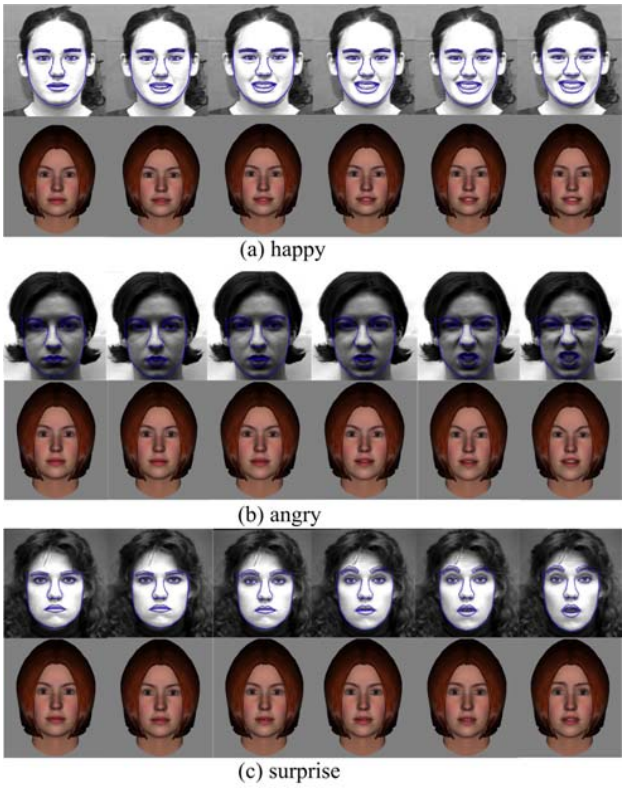

Fig.4 original and synthesized face images

Fig.4 shows some examples of the synthesized face image sequences from the estimated optimal FAP parameters. One can notice that the images not only show well the expressions of different emotions, but also follow the dynamic process of the emotions from neutral to exaggerate. To evaluate how well the synthesized facial animations express the emotions, we listed the 30 synthesized 3D facial animations, as well as the 2D facial animations constructed from the original face image sequences, in a random order, and asked 12 students to recognize their emotions. The results are shown in Table II. One can notice that for happy and sad, most of the synthesized facial animations show correctly the emotions. For fear, angry and surprise, the recognition rates are relatively low. However, this keeps coordinate with the

emotion recognition results on the original 2D facial animations.

TABLE II
EMOTION RECOGNITION RATE OF THE SYNTHESIZED FACIAL ANIMATIONS

Emotion	happy	sad	fear	angry	disgust	surprise
Original 2D	97.1%	94.2%	71.4%	74.2%	77.1%	82.8%
Synthesized 3D	96.67%	90%	55%	70%	75%	61.67%

The 12 students were also asked to perform the subjective evaluation on the synthesized facial animations from the optimally estimated FAPs, as well as from the extracted FAPs of the original face images. The subjects have been asked to assign scores on a five point scale: 1 (bad), 2 (poor), 3 (fair), 4 (good) and 5 (excellent), according to the naturalness and the ability of expressing corresponding emotions. The results are shown in Table III. One can notice that for most of the emotions, the facial animations using the optimally estimated FAPs get better comments. This is mainly due to the fact that the extracted FAPs from the original face images are very sensitive to the tracked feature points, the jerky in the temporal dynamics of the FAPs, as shown in Fig.3, causes jerky of the synthesized facial animations, which thereby influences the naturalness of the animations. On the contrary, since the optimally estimated FAPs are a weighted sum of the mean values of the current state (see Eq. (6)), they are smoother than the extracted FAPs. Therefore, the constructed facial animations are more natural than those from the extracted FAPs.

TABLE III
SUBJECTIVE EVALUATION OF THE SYNTHESIZED FACIAL ANIMATIONS

Emotion	happy	sad	fear	angry	disgust	surprise
Extracted FAPs	3.41	3.17	3.43	3.77	3.56	3.63
Estimated FAPs	3.88	3.58	3.19	3.83	3.64	3.63

VII. CONCLUSIONS AND FUTURE WORK

In this paper, we proposed a DBN based MPEG-4 compliant 3D facial animation synthesis method driven by the continuous emotion space. For each emotion, a two stream state synchronous DBN (SS_DBN) model is trained, with the labeled (Evaluation, Activation) sequences, as well as the extracted FAP parameters of the face image sequences. Once the parameters of the SS_DBN models are trained, given an input (Evaluation, Activation) sequence in the 2D continuous emotional space, the optimal FAP parameters can be estimated via the maximum likelihood estimation criterion, which are then used to construct the MPEG-4 compliant 3D facial animations. Emotion recognition results, as well as subjective evaluations on the constructed facial animations, show that the proposed approach can effectively and efficiently synthesize expressive emotions, as well as follow the dynamic process of the emotions from neutral to exaggerate. In our future work, we will expand the experiments on the audio-visual SEMAINE database, which not only is labeled with both discrete emotions and continuous emotion dimensions, but also has more naturalistic emotions compared to the acted Cohn-Kanade database.

ACKNOWLEDGMENT

This work is supported within the framework of the National Natural Science Foundation of China (61273265), the Shaanxi Provincial Key International Cooperation Project (2011KW-04), the LIAMA-CAVSA project, the EU FP7 project ALIZ-E (grant 248116), and the VUB-HOA CaDE project.

REFERENCES

[1] C. Pelachaud, "Modeling multimodal expression of emotion in a virtual agent," *Philosoph. Trans. Roy. Soc. B Biol. Sci. B*, vol. 364, pp. 3539-3548, 2009.

[2] Plutchik, R., *The Psychology and Biology of Emotion*.Harper Collinns, New York, 1994.

[3] Z. Y. Wu, S. Zhang, L. H. Cai, and H. M. Meng, "Real-time synthesis of Chinese visual speech and facial expressions using MPEG-4 FAP features in a three-dimensional avatar," in *Proc. Int. Conf. Spoken Lang. Process.*, pp. 1802-1805, 2006.

[4] Balci, K., "Xface: MPEG-4 based Open Source Toolkit for 3D Facial Animation", *Proc. Advance Visual Interfaces*, pp. 399-402, 2004.

[5] Lyons, M.J., Akamatsu, S. et al, "Automatic Classification of Single Facial Images, " *IEEE Trans on Pattern Analysis and Machine Intelligence*, pp. 1357-1362, 1999.

[6] J. Cassell, V. Pelachuad, N. Badler, M. Steedman, B. Achorn, T. Becket, B. Douville, S. Prevost, and M. Stone. "ANIMATED CONVERSATION: Rule-based Generation of Facial Expression, Gesture & Spoken Intonation for Multiple Conversational Agents," In *Proceeding of SIGGRAPH' 94*, 1994.

[7] P. Ekman, and W. friesen. Manual for the Facial Action Coding System. *Consulting Psychologists Press*,1978.

[8] E. Costantini, F. Pianesi, and P. Cosi. Evaluation of Synthetic Faces: Human Recognition of Emotional Facial Displays. In E. Andrè, L. Dybkiaer, W. Minker, and P. Heisterkamp, editors, *Affective Dialogue Systems ADS '04, Springer-Verlag*, 2004.

[9] T.D. Bui, D. Heylen, M. Poel, and A. Nijholt, "Generation of Facial Expressions from Emotion Using a Fuzzy Rule Based System," In *Proceedings of the 14th Australian Joint Conference on Artificial Intelligence* (AI 2001), Adelaide, Australia, December 2001.

[10] Mana, N., and Pianesi, F."HMM-based Synthesis of Emotional Facial Expressions during Speech in Synthetic Talking Heads," *Proceedings of ICMI2006*, Banff, Alberta, Canada, November 2006.

[11] Y. Zhang and Q. Ji, "Active and dynamic information fusion for facial expression understanding from image sequence," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 5, pp. 699-714, May 2005.

[12] A. Mehrabian, "Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament," Current Psychol.: *Development., Learn., Personal., Soc.*, vol. 14, pp. 261-292, 1996.

[13] Albrecht I., Schroder M., Haber J., Seidel H., "Mixed feelings: Expression of non-basic emotions in a muscle-based talking

head," *Special issue of Journal of Virtual Reality on "Language, Speech & Gesture"*, to appear.

[14] H. Boukricha et al., "Pleasure-arousal-dominance driven facial expression simulation," in *Proc. ACII Workshops*, pp. 1-7, 2009.

[15] J. Jia et al., "Emotional audio-visual speech synthesis based on pad," IEEE Trans. on Audio, Speech, & Language Processing, pp. 570-582, 2010.

[16] Arya, A., DiPaola, S., Parush, A., "Perceptually valid facial expressions for character-based applications." *International Journal of Computer Games Technol-ogy*, 2009.

[17] Raouzaiou, A., Tsapatsoulis, N., Karpouzis, K., & Kollias, S.,"Parameterized facial expression synthesis based on MPEG-4,"*EURASIP Journal on Applied Signal Processing*, pp. 1021–1038, 2002.

[18] Kanade, T., Cohn, J. F., & Tian, Y., "Comprehensive database for facial expression analysis, " *Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition (FG'00)*, Grenoble, France, pp. 46-53,2000.

[19] Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., Schröder, M., "'FEELTRACE': An instrument for recording perceived emotion in real time," In: *Douglas-Cowie, E., Cowie, R., Schrder, M. (eds.) ISCA Workshop on Speech and Emotion: A Conceptual Framework for Research, Belfast*, pp. 19-24 ,2000.

[20] Iqor S. Pandzic, Robert Forchheimer, *MPEG-4 Facial Animation: The Standard, Implementation and Applications*, John Wiley & Sons, Inc, New York, NY, 2003.

[21] Hou, Y., Sahli, H., Ravyse, I., Zhang, Y., Zhao, R., "Robust Shape Based Head Tracking". *Proc of the Advanced Concepts for Intelligent Vision Systems*. LNCS, pp. 340-351, 2007.

[22] Bilmes, J., Zweig, G., "The Graphical Models Toolkit: An Open Source Software System for Speech and Time Series Processing". *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, pp. 3916-3919, 2002.

[23] S. Zhang , Z.Y. Wu , H. M. Meng , L.H. Cai, " Facial Expression Synthesis Using PAD Emotional Parameters for a Chinese Expressive Avatar," *Proceedings of the 2nd international conference on Affective Computing and Intelligent Interaction*, September , 2007

[24] Balci, K., "Xface: MPEG-4 based Open Source Toolkit for 3D Facial Animation", Proc. *Advance Visual Interfaces*, pp. 399-402, 2004.