SYNTHESIS AND CONTROL OF HIGH RESOLUTION FACIAL EXPRESSIONS FOR VISUAL INTERACTIONS

Chan-Su Lee, Ahmed Elgammal, and Dimitris Metaxas

Rutgers University
Computer Science Department
New Brunswick, NJ 08854, USA

ABSTRACT

The synthesis of facial expression with control of intensity and personal styles is important in intelligent and affective human-computer interaction, especially in face-to-face interaction between human and intelligent agent. We present a facial expression animation system that facilitates control of expressiveness and style. We learn a decomposable generative model for the nonlinear deformation of facial expressions by analyzing the mapping space between low dimensional embedded representation and high resolution tracking data. Bilinear analysis of the mapping space provides a compact representation of the nonlinear generative model for facial expressions. The decomposition allows synthesis of new facial expressions by control of geometry and expression style. The generative model provides control of expressiveness preserving nonlinear deformation in the expressions with simple parameters and allows synthesis of stylized facial geometry. In addition, we can directly extract the MPEG-4 Facial Animation Parameters (FAPs) from the synthesized data, which allows using any animation engine that supports FAPs to animate new synthesized expressions.

1. INTRODUCTION

In intelligent and affective computing, computers need abilities not only to recognize human emotion but also to express emotions to convey affections to the human [1]. Conversational agents are frequently used in affective computing to express emotions even though computers that have very different bodies can be used to display emotions [2]. Facial expressions, however, are the main communication channel of emotions in face-to-face interactions. We present a facial animation system for effective emotion communications for intelligent agents.

It is important to generate subtle details of facial expression that convey personality and intensity in facial expressions. We want to generate facial expressions that give the appearance with different expressiveness and with different personality in emotion expressions. Our research is related to emotional appearance and multiple levels of emotion generation in affective computing [3]. The challenging aspects in realistic facial expression synthesis are capturing details of facial expression including small nuances [4] in different persons and generating new stylized expressions with the control of expressiveness. We learn nonlinear mapping of facial motions in order to capture nonlinear deformations in facial expressions. The decomposition of the mapping space using bilinear model allows us to parameterize subtle expression characteristics in different people with realistic facial expression synthesis. We present the representation of facial expression using a generative model in addition to the high resolution tracking in Sec. 2.

We provide control of intensity and personality in synthesized facial expressions. Exaggerations of differences between an expressive face and a neutral face enhances intensity of the given expression [5]. The exaggeration of differences in geometry of a person face from standard face geometry synthesizes stylized face. We also generate new stylized expression by combination of existing styles. The scaling of facial motion provides control of expressiveness, intensity of facial expressions. In Sec. 3.1 and 3.2, we explain the synthesis of facial expression with expressiveness control and stylized facial geometry and the estimation of the animation parameter using the proposed generative model.

From synthesized facial animation, we extract MPEG-4 facial animation parameters (FAPs), low resolution animation parameters to drive facial animation in other animation engine for conversational agents. As we have very high resolution tracking data for facial expression and synthesis of the facial nodal points, we can extract FAPs after identifying feature points for parameter extractions from a generic model used for tracking. Extracted FAPs are used to synthesize facial animation in animiation softwares that support MPEG-4 FAP format. Synthesis results of facial expressions using FAPs are presented in Sec. 3.3.

Related Works: Synthesis of facial expressions is preformed by modeling facial expressions using generative models and controlling model parameters. Researchers have used linear models (PCA [6]) and variations such as bilinear models [7, 8] and multilinear tensor models for facial expression analysis

and synthesis [8, **?**]. However, a major limitation of such models is modeling the facial expression in linear subspaces. In this paper, we analyze expression in a nonlinear mapping space and synthesize new facial expressions based on the decomposition of the mapping space.

Exaggerations of face were investigated by Brennan [11] in computer graphics to develop a caricature generator using user's line drawing input with manual correspondence. Exaggeration was controlled by scaling difference in the corresponding points. There are few works related to 3D facial caricature and facial expression exaggeration. We provide stylized facial expressions which support caricature of face geometry and exaggeration of 3D facial motions.

2. GENERATIVE MODELS FOR HIGH RESOLUTION FACIAL EXPRESSION SYNTHESIS

We achieve high resolution tracking of facial expressions from dense clouds of 3D range data using harmonic maps [14]. Tracking data provides an efficient non-rigid 3D motion tracking with correspondences between different frames of the same sequences as well as between different sequences.

2.1. Learning nonlinear mapping space of facial expressions

Facial motion can be described by the displacements of 3D facial nodal points of the general face geometry used for tracking. As a result of high resolution tracking with one-to-one intra-frame correspondence, we can represent facial expressions by motion vectors for the vertices of a generic face mesh. Let $\boldsymbol{v}_t \in R^{3N\times 1}$ be locations of 3D points at time instance t representing N facial nodal points in a 3-dimensional space, where N is the number of nodal points in a dense generic facial model. The trajectory of the 3D nodal points is the combination of rigid head motion and facial motion, which can be described as

$$\boldsymbol{v}_t = T_{\boldsymbol{\alpha}_t}\boldsymbol{y}_t = T_{\boldsymbol{\alpha}_t}\left(\boldsymbol{y}_0 + \boldsymbol{m}_t\right) = T_{\boldsymbol{\alpha}_t}(\boldsymbol{g} + \boldsymbol{m}_t)\,, \quad (1)$$

where $T_{\boldsymbol{\alpha}_t}$ is the head motion at time t, $\boldsymbol{y}_t$ is the $3 \times N$ face nodal point locations at time t in face centered coordinate and $\boldsymbol{y}_0 = \boldsymbol{g}$ is the facial geometry at the initial frames. We assume that the captured facial expression starts from a neutral face. The global rigid transformation parameter $\boldsymbol{\alpha}$ comes from the tracking results.

Tracking data are collected from multiple people for learning stylized facial expression models. The displacement in local coordinate of every facial nodal point from the tracking data of expression style s can be described as

$$\boldsymbol{v}_t^s = T_{\boldsymbol{\alpha}^s}(\boldsymbol{g}^s + \boldsymbol{m}_t^s)\,, \quad (2)$$

where $T_{\boldsymbol{\alpha}^s}$ is the global transformation by head motion which depends on expression style and type, $\boldsymbol{g}^s$ is the facial geometry of each person, and $\boldsymbol{m}_t^s$ is the facial expression motion.

The main problem in facial expression animation is how to model and control the facial expression motion, $\boldsymbol{m}_t^s$, and facial geometry $\boldsymbol{g}^s$. Both of them depend on the person. The facial motion undergoes nonlinear deformations and it is of high dimension. We derive a low dimensional representation for facial motion using conceptual manifold embedding. Unit circle is used in the embedding of facial expression which is cyclic in the sense that the expression changes from neutral expression → target expression → neutral expression in the tracking data set. The conceptual manifold is homeomorphic to actual data-driven manifold using nonlinear dimensionality reduction like LLE (locally linear embedding) [15] that finds intrinsic configuration representation in low dimensional space [16].

Given a set of distinctive facial motion sequence $\boldsymbol{M}^s = [\boldsymbol{m}_1^s\boldsymbol{m}_2^s\cdots\boldsymbol{m}_{N_s}^s]^T$ and its embedding $\boldsymbol{X}^s = [\boldsymbol{x}_1^s\boldsymbol{x}_2^s\cdots\boldsymbol{x}_{N_s}^s]^T$, we can learn nonlinear mapping function $f^s(\boldsymbol{x})$ that satisfies $f^s(\boldsymbol{x}_i) = \boldsymbol{m}_i^s$, $i = 1\cdots N_s$, where N_s is the number of captured motion frame for style s. Generalized radial basis function (GRBF) interpolation [17] is used to learn mapping in the form $f^s(x) = \boldsymbol{B}^s\psi(x)$ where each row in the matrix $\boldsymbol{B}$ represents the interpolation coefficients for corresponding element in the input. i.e., we have d simultaneous interpolation functions each from $2D$ to $1D$. The mapping coefficients can be obtained by solving the linear system

$$[\boldsymbol{m}_1^s\cdots\boldsymbol{m}_{N_s}^s] = \boldsymbol{B}^s[\psi(\boldsymbol{x}_1^s)\cdots\psi(\boldsymbol{x}_{N_s}^s)] \quad (3)$$

The mapping function contains all the information to generate new interplated motion for given $\boldsymbol{x}_t$. For a given kernel $\psi(x)$, the matrix $\boldsymbol{B}^s$ captures the facial motion characteristics for expression style s. As a result, the facial expression of person style s can be represented by

$$\boldsymbol{v}_t^s = T_{\alpha_s}(\boldsymbol{g}^s + \boldsymbol{B}^s\psi(\boldsymbol{x}_t)). \quad (4)$$

However, this model requires to high dimensional parameters $\boldsymbol{g}^s$ and $\boldsymbol{B}^s$ for each person to generate a new facial expression.

2.2. Decomposition of facial geometry and facial motion

We achieve compact and orthogonal representation for face geometry and facial motion mapping space using bilinear analysis. We can represent matrix $\boldsymbol{B}^s$ as a vector $\boldsymbol{b}^s$ by column stacking. We collect mapping vector $\boldsymbol{b}^s$ as

$$\boldsymbol{F} = [\boldsymbol{b}^1\boldsymbol{b}^2\cdots\boldsymbol{b}^{N_s}]\,, \quad (5)$$

where N_s is the number of facial expression styles. By applying an asymmetric bilinear model [7], we can represent geometry based on geometry basis $\boldsymbol{E}$ and geometry style vector $\boldsymbol{s}_b^s$ as follows:

$$\boldsymbol{F}^s = \boldsymbol{E}\boldsymbol{s}_b^s\,. \quad (6)$$

We apply the bilinear analysis for facial geometry and achieve representation of person geometry based on geometry basis

$\boldsymbol{D}$ and geometry style vector $\boldsymbol{s}_g^s$ as follows:

$$\boldsymbol{C}^s = \boldsymbol{D}\boldsymbol{s}_g^s \,, \tag{7}$$

where the dimension of the style vector $\boldsymbol{s}_g^s$, a compact representation of person geometry, is $N_s (<< 3N)$.

Now, the decomposable nonlinear generative model can be expressed as

$$\boldsymbol{v}^s(t) = T_{\boldsymbol{\alpha}_t^s}\left(\boldsymbol{D}^g\boldsymbol{s}_g^s + unstacking(\boldsymbol{E}\boldsymbol{s}_b^s)\psi(\boldsymbol{x}(t))\right), \tag{8}$$

where $unstacking(\cdot)$ means converting vectorized representation to original matrix by unstacking column. The generative model captures nonlinearity in the mapping space and provides compact control parameters for expression style and expression geometry variations in facial expression synthesis. In addition, it supports generation of stylized geometry and facial expression.

3. SYNTHESIS OF STYLIZED AND EXPRESSIVE FACIAL EXPRESSIONS

New facial expressions can be generated by new geometry style vectors and new facial motion style vectors. Linear weighting of the existing style vector can be used to generate new style vector for synthesis.

$$\begin{aligned} \boldsymbol{s}_g^{new} &= \alpha_1\boldsymbol{s}_g^1 + \alpha_2\boldsymbol{s}_g^2 + \cdots + \alpha_{N_e}\boldsymbol{s}_g^{N_s} \,, \\ \boldsymbol{s}_b^{new} &= \beta_1\boldsymbol{s}_b^1 + \beta_2\boldsymbol{s}_b^2 + \cdots + \beta_{N_s}\boldsymbol{s}_b^{N_s} \,, \end{aligned} \tag{9}$$

where $\sum_i \alpha_i = 1$, and $\sum_j \beta_j = 1$. α_i controls the weight for geometry style $\boldsymbol{s}_g^i$, and β_j specify the weight for facial motion style $\boldsymbol{s}_b^j$. New facial expression can be generated by using new styles as

$$\boldsymbol{v}^{new}(t) = T_{\boldsymbol{\alpha}_t^{new}}\left(\boldsymbol{D}^g\boldsymbol{s}_g^{new} + unstacking(\boldsymbol{E}\boldsymbol{s}_b^{new})\psi(\boldsymbol{x}(t))\right). \tag{10}$$

We can add additional parameters to synthesize more stylized geometry and to control expressiveness.

3.1. Stylized facial geometry synthesis

New stylized facial geometry can be synthesized using mean geometry style. Stylized face geometry exaggerates facial features that are different from standard face, or average face. As we get tracking data with correspondence between different subjects, the mean facial geometry can be the arithmetic mean of individuals' face geometry. The same result can be achieved using equal weighting of all the style vectors. We can represent using mean style $\bar{s}$ and scale factor γ as

$$g^{exaggerated} = D^g(\gamma(s_g^s - \bar{s}_g) + \bar{s}_g) \,. \tag{11}$$

The parameter γ controls the amount of exaggeration based on difference from the average face geometry. When $\gamma = 1$, the new face is the same face as the original face. When $\gamma < 1$, the new face closes to standard face and the same as average face at $\gamma = 0$. Usually, caricatured faces are generated using $\gamma > 1$. Fig. 1 shows mean geometry, and style exaggerated geometries.

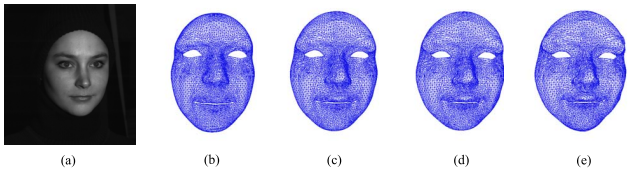


Fig. 1. Original image and its 3D caricatures: Col.(a): Original images Col. (b): $\gamma = 0.0$ (mean geometry), Col. (c): $\gamma = 0.7$, Col. (d): $\gamma = 1.0$, Col. (e): $\gamma = 1.3$.

3.2. Intensity control in facial expressions

We can control expressiveness in the synthesized facial expression by scaling the facial motion. We can extend the generative model using motion scaling parameter τ.

$$\boldsymbol{m}_t^{exaggerated} = \tau(unstacking(\boldsymbol{E}\boldsymbol{s}_b^{new})\psi(\boldsymbol{x}(t))) \,, \tag{12}$$

The scaling of the motion vector affects different strength or feeling of the expression. Reducing the facial motion scale parameter τ causes milder or weak expression than the original ones and stronger expression when we increase scaling parameter. The overall generative model is combination of stylized facial geometry and facial motion with expressiveness control. Fig. 2 show examples of scaling effects, which shows different expressiveness of facial motion according to scale parameters in smile expression sequence. Facial expressions in different intensities are generated with preserving details in the expression.

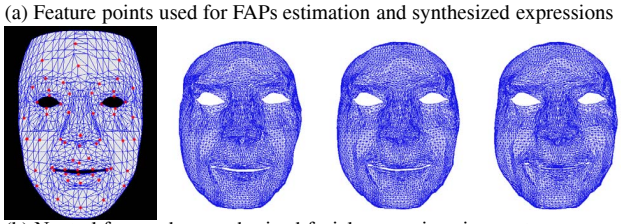
Fig. 2. Intensity control by scaling factors in facial expressions: Row: τ changed by 0.5, 1.0, 1.25, 1.75 from top row to button row. Column: Sampled expression sequence(neutral → target)

3.3. Extraction of FAPs and animation in intelligent agents

MPEG-4 facial animation parameters (FAPs) are part of the specifications of the international standard for efficient coding of shape and animation of human face and bodies. FAPs are used not only for the web and mobile applications but also emotion recognition and talking head human computer interface [18]. Computer vision techniques are used to extract FAP parameters from video. Optical marker based motion capture systems are used to get accurate parameters as in [19]. Our high resolution tracking of facial motion by generic model with correspondence allows accurate extraction of FAPs without markers.

As we have high resolution tracking of 3D facial nodal points with correspondence between frames and in different sequences, we can estimate the facial animation parameters (FAPs) if we once identify the nodal points corresponding to FAP nodes. Our generic facial geometry model does not create for FAP parameters. However, it has 8K 3D nodal points in high resolution tracking and 1K in low resolution tracking. We can easily identify nodal points to calculate FAPs. In Fig. 3, the first colume shows identified nodal points in neutral face. In the following columes, it shows synthesized facial expressions (a) and corrresponding re-synthesis results (b) of expressions using extracted FAPs from the synthesis data. We used *Visage Technologies software* supplied by Visage Technologies AB to convert FAPs into binary format and to generate facial expressions in agents.

(a) Feature points used for FAPs estimation and synthesized expressions



(b) Neutral face and re-synthesized facial expressions in agents

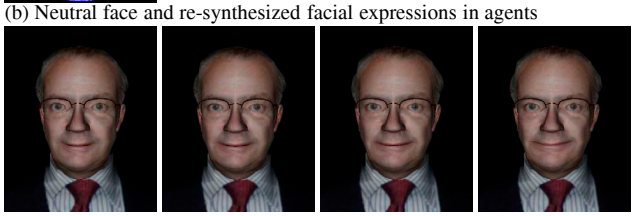


Fig. 3. Feature points and re-synthesis of facial expressions

4. CONCLUSION AND FUTURE WORK

We presented a facial expression synthesis system using a nonlinear generative model from tracking data of high resolution 3D nodal points with correspondences between frames and in different persons. The generative model can synthesize, not only personalized facial expressions including subtle expressions, but also expressiveness controlled ones. In addition, we can extract FAPs from the synthesized data and can generate facial expression by any animation software supporting MPEG-4 FAPs. We plan to perform evaluation of our system in human-intelligent agent interaction situation to see the effect of expressiveness and personalization in intelligent agent appearance.

Acknowledgement: The facial expression tracking data was made available by Dimitris Samaras at SUNY Stony Brook. This research is partially funded by NSF award IIS-0328991.

5. REFERENCES

[1] Rosalind W. Picard, "Affective computing: Challenges," *Int. Journal of Human-Computer Studies*, vol. 59, no. 1–2, pp. 55–64, 2003.

[2] Karen K. Liu and Rosalind W. Picard, "Subtle expressivity in a robotic computer," in *Proc. of the CHI 2003 Workshop on Subtle Expressivity for Chnaracters and Robots*, 2003.

[3] Rosalind W. Picard, "What does it mean for a computer to have emotions?," Tech. Rep. 534, MIT Media La, 2001.

[4] Bernadette Kiss, Balazs Benedek, Gabor Szijarto, and Barnabas Takacs, "Closed loop dialog model of face-to-face communication with a photo-real virtual human," in *Visual Communications and Image Processing*, 2004, pp. 765–772.

[5] Andrew J. Calder, Duncan Rowland, Andrew W. Young, Ian Nimmo-Smith, Jill Keane, and David I. Pettett, "Caricaturing facial expressions," *Gognition*, vol. 76, pp. 105–146, 2000.

[6] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3d faces," in *SIGGRAPH'99*, 1999, pp. 187–194.

[7] Joshua B. Tenenbaum and William T. Freeman, "Separating style and content with biliear models," *Neural Computation*, vol. 12, pp. 1247–1283, 2000.

[8] E. S. Chunang, H. Deshpande, and C. Bergler, "Facial expression space learning," in *Pacific Graphics*, 2002, pp. 68–76.

[9] T. Ezzat, G. Geiger, and T. Poggio, "Trainable videorealistic speech animation," in *SIGGRAPH*, 2002, pp. 388–398.

[10] Daniel Vlasic, Matthew Brand, Hanspeter Pfister, and Javan Popovic, "Multilinear models for face synthesis," in *SIGGRAPH Sketches*, 2004.

[11] Susan E. Brennan, "Caricature generator: The dynamic exaggeration of faces by computer," *Leonardo*, vol. 18, no. 3, pp. 170–178, 1985.

[12] Rein-Lien Hsu and Anil K. Jain, "Generating discriminating cartoon faces using interacting snakes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 11, 2003.

[13] Lin Liang, Hong Chen, Ying-Qing Xu, and Heung-Yeung Shum, "Example-based caricature generation with exaggeration," in *Proceedings of Pacific Conference on Computer Graphics and Applications (PG'02)*, 2002, pp. 386–403.

[14] Yang Wang, Mohit Gupta, Song Zhang, Sen Wang, Xianfeng Gu, Dimitris Samaras, and Peisen Huang, "High resolution tracking of non-rigid 3d motion of densely sampled data using harmonic maps," in *Proc. in ICCV*, 2005, pp. 388–395.

[15] Sam Roweis and Lawrence Saul, "Nonlinear dimensionality reduction by locally linar embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[16] Ahmed. Elgammal and Chan-Su Lee, "Separating style and content on a nonlinear manifold," in *CVPR*, 2004, pp. 478–485.

[17] Tomaso Poggio and Fredrico Girosi, "Networks for approximation and learning," *Proceedings of the IEEE*, vol. 78, no. 9, pp. 1481–1497, 1990.

[18] Igor S. Pandzic and Robert Forchheimer, Eds., *MPEG-4 Facial Animation: The Standard, Implementation and Applications*, John Wiley, 2002.

[19] Sumedha Kshirsagar, Stephane Garchery, Gael Sannier, and Nadia Magnenat-Thalmann, "Synthetic faces: Analysis and applications," *Int. J. Imaging Syst. Technol.*, pp. 65–73, 2003.