A Facial Animation System for Generating Complex Expressions

Chunyu BAO* Ee Ping ONG† Arthur NISWAR‡ Zhiyong HUANG§

Institute for Infocomm Research, Singapore
* E-mail: cbao@i2r.a-star.edu.sg
† E-mail: epong@i2r.a-star.edu.sg
‡ E-mail: aniswar@i2r.a-star.edu.sg
§ E-mail: zyhuang@i2r.a-star.edu.sg

***Abstract*—This paper presents a novel expressive facial animation system based on a motion captured data stored in a database. Unlike common data-driven facial animation systems which only use mo-cap data of one expression at a time, our system can use data of more than one expression. Thus, the system can create a complex expression when the virtual human is talking. Using our method, the resulting facial animation can be a combination of the expressions from different faces in databases. In this way, complex expressive animation can be generated, even some extreme ones that are difficult for an actor/actress, e.g., a happy expression on the left side of the face that smoothly transits to the sad on the right. This paper will describe the system and show some experimental results.**

I. Introduction

Expressive facial animation has a wide usage in the entertainment industry, human computer interaction and so on. From decades before till now, various facial animation methods have been developed. The early research focused on physically-based modeling techniques [1] [2] [3] [4] [5] [6], which generate facial expressions by simulating muscle and skin movement, such as Facial Action Coding System (FACS) [2]. FACS decomposes the facial emotion into small Action Units(AUs) based on facial anatomy and uses the movement of some specific AUs to generate different expressions.

Recently, performance-driven techniques which produce facial movement based on the facial motion data of real people have been widely explored [7] [8] [9] [10] [11] [12] [13] [14] [15] [16] [17] [18] [19]. Compared to physically-based modeling, performance-driven techniques can potentially generate more realistic and more natural emotion. Using performance-driven techniques, a large face expressions database needs to be collected.

There are two categories of performance-driven techniques. One category is called phoneme-based methods, which connect phoneme segments directly from facial motion database [11] [20] [21]. Kshirsagar and Thalmann [11] presented a method in which motion-captured data are categorized and divided into small parts based on syllable motions and new speech animation is synthesized by concatenating syllable motions from created database. In [21] [22] [23], the search algorithms for created database were improved in order to combine longer phoneme or syllable sequences.

Another category is called data-driven facial animation approaches, which learn statistical model from motion-captured data. Chuang et al. [12] [13] learned a mapping from neutral emotion to other emotions using bilinear models. To synthesize an expressive facial animation, the facial animation with neural emotion is synthesized first and emotional animation is created through mapping. Cao et al. [7] set up a mapping from neutral emotion to other emotions through a Radial Basis Function(RBF) and this mapping is used to transfer neutral talk to expressive talk. Deng et al. [8] aligned expressive motion data with neutral motion data by phoneme-based time warping and extract pure expressive motion signals by subtracting neutral motion from expressive data. These pure expressive motion signals are used to generate expressive facial animation from neutral ones.

The techniques used in our system belong to the second category, which set up statistical model from motion-captured data. Unlike common data-driven facial animation systems which only use motion data of one expression at one time, our system can use data of more than one expression. Sometimes, for some special effects, there may be needs to show different expressions simultaneously on a face. For example, the movie star Jim Carey showed a very complex expression on his face in his famous movie "The Mask". In order to synthesize such special effects, we design the system that can create more kinds of expressions on the human face when he/she is talking. For example, we can have expressions of happy on left side of the face but disgust or neutral on right side of the face. Such effects cannot be created from database using purely performance-based face emotion synthesis approaches, as it is very difficult for actor/actress to produce such expressions and be captured. Here, we explore using a combinational approach for complex face expression synthesis to solve the above-mention problem where we are able to synthesis different expressions on both the left/upper part of the face with the right/lower part of the face coupled with lips synchronization for synthesized speech. One problem in our solution is creating a seamless transition between different face parts so that the expressions on the synthesized face still looks plausible albeit

Fig. 1. The snapshots of the captured actress with 103 markers on her face.

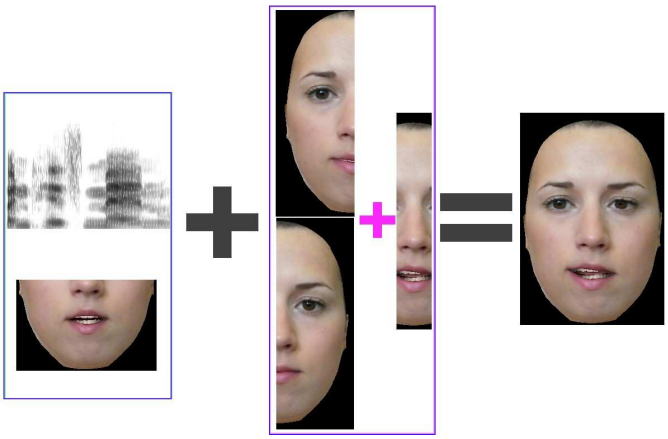


Fig. 2. The expressive facial animation is created by two subsystems: visual speech synthesis (left) and complex expression synthesis (middle).

the difficulty where only somebody with special talent is able to produce. In this paper, we present a complex expression synthesis system for such purpose.

The rest of the paper is organized as follows. In Section II, we give a brief introduction of facial motion data collection and processing. In Section III, the structure of our complex expressive facial animation system is given. The process and method of visual speech synthesis and complex expression synthesis are presented in Section IV and Section V respectively. Results are shown and discussed in Section VI, followed by conclusions in Section VII.

II. Facial Motion Data Collection and Processing

To acquire high quality data, a VICON motion capture system was used to capture the motions of a professional actress, who had 103 markers on her face (Figure 1). The actress was arranged to speak an exquisitely designed corpus seven times, with different emotions, including neutral, happy, sad, angry, fear, surprised, disgusted emotions. The corpus is composed with 539 phoneme-balanced sentences. The facial motion data was captured with a rate of 100 frames per second. The values of some markers were removed because of the tracking error and 92 markers are kept.

After data collection, we normalized the facial motion data. A neutral emotion frame with closed-mouth pose was chosen as a reference frame. The other data was aligned to the reference frame through translating, scaling, and rotation.

We used a speech recognition engine to align recorded audio with corresponding phonemes in the forced-alignment mode. The aligned results are checked one by one to correct some errors. Then the alignment results were used to align each phoneme with its corresponding motion data segments. After that, based on phoneme, we aligned expressive captured data strictly with neutral emotion data through time warping and re-sampling. Pure expressive motion signals was extracted by subtracting neutral motion frames from aligned expressive motion frames.

After that, we removed head motions from captured motion data. Then principle component analysis (PCA) algorithm was used to reduce the dimensionality of these motion capture vectors. We reduced the dimension to 5, covering over 90.1% of variation.

III. System Structure

Our system is composed of two subsystems: visual speech synthesis and complex expression synthesis (Figure 2). The visual speech synthesis subsystem is to produce the speech audio and reconstruct the trajectory of the visemes in the speech. The speech audio is generated by TTS (Text-To-Speech) sytem and the speech animation is based on the visemes. The complex expression synthesis subsystem is used to construct the complex expressive motion of the face.

IV. Visual Speech Synthesis

Speech animation is basically the reconstruction of the visemes (the visual counterpart of phonemes) that constitute the speech. The visemes are represented by their parameters, which are extracted from a synthetic viseme database using PCA [24]. Six parameters are used to describe each English viseme, so each viseme can be represented using a column vector α with 6 components.

The visual speech is synthesized by computing trajectory of α for the whole speech segment, based on the phoneme durations generated by a TTS engine, given the input text. This TTS system also generates the audio signal of the speech. The trajectory of α (which represents the visemes) for the whole speech is then calculated as follows:

- Determine the appropriate viseme for each phoneme generated by the TTS system: one viseme for a monophthong, two visemes for a diphthong, and no viseme for phoneme /h/ (since its visual realization follows the preceding or subsequent phoneme). For speech pause, the neutral head pose with closed mouth is assigned as the corresponding viseme.
- Assign α of the corresponding viseme in each phonemic segment as follows:

  - For a monophthong, set α at the median of the segment.
  - For a diphthong, there are two αs. Set the first α at the first quartile of the segment and the second α at the third quartile.
  - For phoneme /h/, there is no α to set in the segment.
- For the speech pause, α is assigned as follows:
  - For the speech pause in the beginning of the speech, α is set at $t = 0$, the first and the second quartile of the segment.
  - For the speech pause at the end of the speech, α is set at $t = T$ (here T is the speech duration which is generated by the TTS engine), the second and the third quartile of the segment.
  - For the speech pause in the middle of the speech, α is set at the first, second, and the third quartile of the segment.
- Determine the value of α in the remaining time using cubic spline interpolation [25] from the assigned α.

V. Complex Expression Synthesis

Now we describe the steps for complex expression synthesis. Firstly, the expressive motions of the facial markers are generated. After that, the synthesized motions of facial markers are mapped to 3D face model; Finally, joint area is carefully processed.

A. Facial Maker Motion Synthesis

As we mentioned before, pure expression data are extracted from the motion database through phoneme-based time warping. The pure expression data are in high dimension. Thus, the data is further processed by PCA algorithm to reduce to 5 dimensions. the results of PCA shows the expression data are continuous curve [8].

The facial marker motions are synthesized using patch-based sampling algorithm for its real-time efficiency [26]. For a complex expression motion, the different area of face may show different emotions, for instance, a happy expression on the left side of the face and an angry one on the right side of the face. In this case, patch-based sampling algorithm may run multiple times to produce a data set with happy expression and a data set with angry expression.

The process is described as follows:

- Determine expression duration according to the Speech duration T which is generated by the TTS system.
- According to the complex expression from in the input panel, extract the expression types which you want to generate. For example, if the input panel shows happy on the left side and angry on the right, the extracted expression types will be happy and angry.
- Generate each type of expression. The process is described as follows:
  - From the database of the corresponding expression, randomly choose a continuous segment with 30 frame data as the initial frame samples. A continuous segment with a fix number of frames is called a patch. Here a patch has 30 frames.
  - Search the database for patch candidates which match the boundary of the former generated frame samples. To match one patch to another patch, their distances of the boundary zone must be smaller than some tolerance extent. Here the boundary zone has 5 frames and the tolerance extent is chosen as 0.1.
  - Randomly choose one patch from the candidates and attach it to the former generated frame samples.
  - Repeat the last two steps till all the samples are found.
- After the frame samples of all the types of expression is generated, project back frame samples from the feature space to the original motion marker space.
- Synthesize the motion data based on the different expressions of different face area. According to the complex expression from in the input panel, the markers of the whole face are divided into several areas. For each area, only the frame samples belonging to the corresponding expression are kept; the frame samples belonging to other expressions are set to 0. For instance, we have a complex expression with happy on the left side of the face and angry on the right. For the frame samples with happy expression, the motions of the markers on right side of the face are set to 0 while the motions of the markers on the left side of the face and the joint area are kept. It is similar for the frame samples with angry expression, the motions of the markers on the left side of the face are changed to 0 while the motions of the markers on the right and the joint area are unchanged.

Normally, humans usually change their expression intensity from time to time. Thus, a changing-expression curve scheme is used in our simulation. Basically, a changing-expression curve can be any continuous curve in time versus intensity space, and its range is from 0 to 1, where 0 means no expression (neutral) and 1 means maximum expression. The changing-expression curve is used to control the intensity of synthesized complex expression signals.

B. 3D Face Deformation

Our 3D face model is produced from a good quality photograph of a person. Give a photograph with front view or even slanted image, we use the technique of Automatic and Real-time 3D Face Synthesis introduced in [27] to generate a good texture mapped 3D model.

After the motions of facial markers are synthesized, we need to map these motions to the above 3D face model. Each marker point is associated with a location on the mesh of the 3D face model. In section 2, we choose a neutral emotion frame with closed-mouth pose as a reference frame. Some feature points are extracted from this reference frame. Base on these feature points, we set up a mapping from motion data to 3D face model [27]. Thus, the motions of facial marker y can map to the 3D face model through a transformation D. The facial

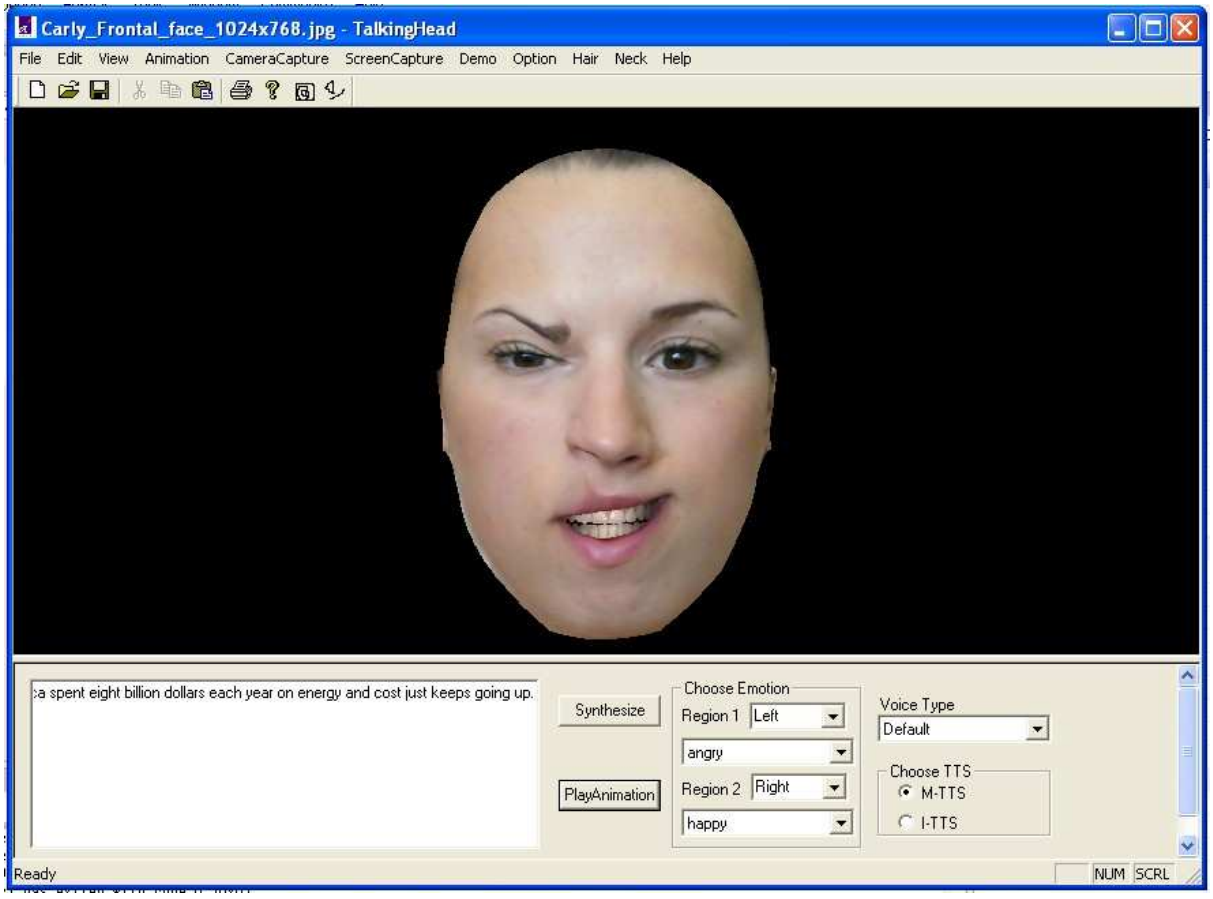


Fig. 3. A snapshot of our running system. The upper part is the facial animation window and lower part is the control panel. The control panel enclose speech-text-input edit control (bottom-left), control buttons (bottom-middle),some option selection boxes (bottom-right).

marker motion M on its corresponding position of the 3D face model can be written as:

$$M = Dy. \tag{1}$$

Here the feature points based mesh deformation method is used to map the motions of the facial markers to other vertices on the 3D face model due to its efficiency. For a target vertex, if its distance to a facial marker is less than a threshold, we say the target vertex is affected by this facial marker. The contribution weights of the facial markers to its affected vertices are defined based on their distances. The contribution weight $W_{k,i}$ of facial marker k to vertex i is written as:

$$W_{k,i} = C \cdot exp(-r_{k,i}/p) \tag{2}$$

where C is a constant value, $r_{k,i}$ is the distances between facial marker k and vertex i and parameter p is experimentally selected. In our experiments, C is set to 1 and parameter p is set to 64. The displacement δ_i of vertex i caused by expression is computed as follows:

$$\delta_i = \frac{\sum_k M_k \cdot W_{k,i}}{\sum_k W_{k,i}} \tag{3}$$

where M_k is the expressive motion of facial marker k.

C. Joint Area Processing

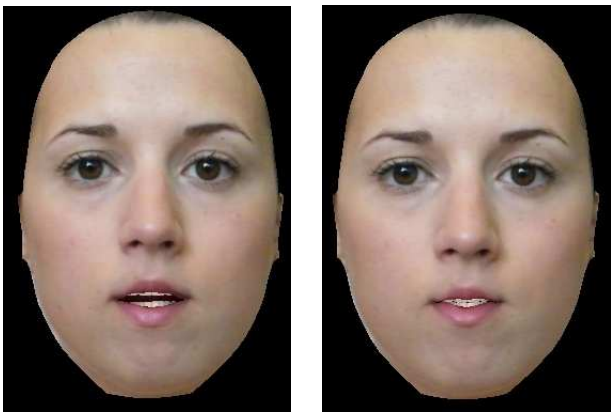
For a complex expression shown on the face, for example, a happy expression on the left side of the face and an angry one on the right, the face is divided into three regions: left region, right region and joint region. The width of joint area is experimentally selected. The displacement of a vertex on the left region, denoted as $\delta_{i,1}$, is computed using only the data set of happy expression; The displacement of a vertex on the right region, say $\delta_{i,2}$, is computed using only the data set of angry expression.

Now the question is how to make the joint area connected continuously and smoothly when a complex expression shows on the face. Here we design an algorithm of linear combination to solve it. For a vertex in the joint area, $\delta_{i,1}$ and $\delta_{i,2}$ are computed respectively using the data sets of happiness and anger. The motion of the vertex in the joint area $\delta_{i,joint}$ is a linear blend of them:

$$\delta_{i,joint} = w \cdot \delta_{i,1} + (1-w) \cdot \delta_{i,2} \tag{4}$$

$$w = (x_{end} - x)/(x_{end} - x_{start}) \tag{5}$$

where x_{start} and x_{end} are the positions of start point and end point of the joint area and x is the position of vertex i.



(a) (b)

Fig. 4. Snapshots of the visual speech (neutral expression) synthesized by our system

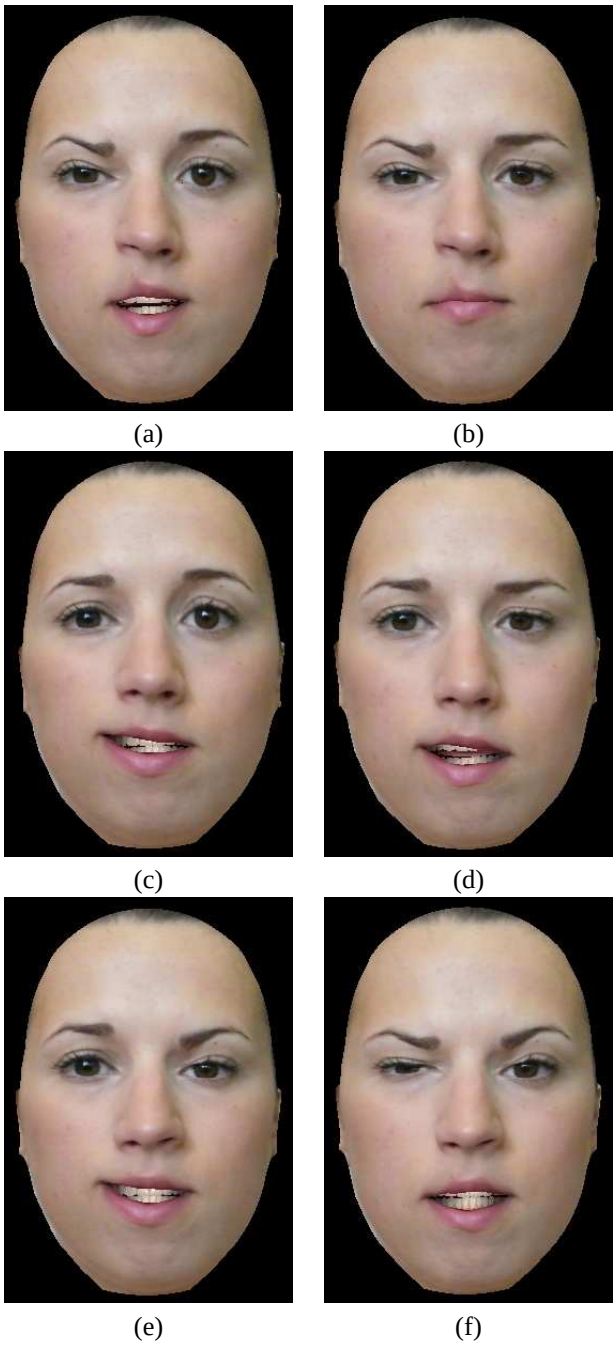
After that, the resulting complex facial expression is combined with visual speech face movement to create the final results.

VI. Results

The complex expressive facial animation system is developed using VC++ 8.0 that runs on the MS Windows XP system. Figure 3 shows a snapshot of the running system. The system includes two parts: the upper part is the animation window and the lower part is the control panel. The animation window is used to show expressive facial animation. The left side of the control panel is the text-input edit control where users can input the speech contents. The right side of the control panel has some selection buttons where users can select the area and expressions for region 1 and region 2, voice type and the TTS engine which is used to synthesize the speech. After speech contents and the complex expression are decided, the animation can be synthesized and played by the middle buttons.

We tested the effectiveness of our complex expressive facial animation system with a number of experiments. Figure 4 shows two snapshots of the visual speech (neutral expression) synthesized by our system.

The snapshots of complex expressive speech animation with different expressions on the left and right face are shown in Figure 5. By using the different data sets from the motion database, we synthesized complex expression with different emotions on the left face region and the right face region respectively: 5(a) shows angry expression on the left face region and neutral expression on the right face region; 5(b) shows angry expression on the left and sad expression on the right; 5(c) shows happy expression on the left and neutral expression on the right; 5(d) shows happy expression on the left and sad expression on the right; 5(e) shows happy expression on the left and angry expression on the right; and 5(f) shows disgust expression on the left and angry expression on the right. From these figures, we can see that mouth



(a) (b) (c) (d) (e) (f)

Fig. 5. Snapshots of the results: the complex facial expression by using the data sets of (a) angry and neutral, (b) angry and sad, (c) happy and neutral, (d) happy and sad, (e) happy and angry, and (f) disgust and angry expressions respectively on the left face and the right face.

may be non-symmetrical on the left side and the right side, because mouth will move up or down when showing different expressions.

The snapshots of complex expressive speech animation with different expressions on the upper and lower part of the face are shown in Figure 6: 6(a) shows angry expression on the upper face region and happy expression on the lower face

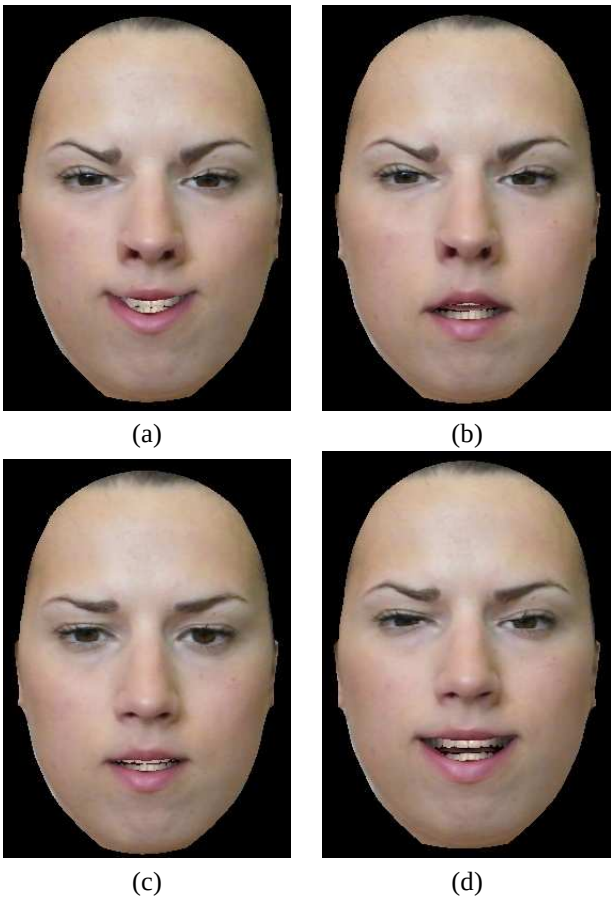


Fig. 6. Snapshots of the results 2: the complex facial expression by using the data sets of (a) angry and happy, (b) angry and neutral, (c) sad and neutral, and (d) disgust and happy expressions respectively on the upper face and the lower face.

region; 6(b) shows angry expression on the upper face and neutral expression on the lower face; 6(c) shows sad expression on the upper face and neutral expression on the lower face; 6(d) shows disgust expression on the upper face and happy expression on the lower face. We can see that the effects of expressions on the upper face are mainly reflected around eyes and the effects of expressions on the lower face are mainly reflected around mouth.

The results demonstrate the ability of the system to generate a high quality rendered animation of complex expressive facial motion.

VII. Conclusion

In this paper, we presented a novel complex expressive facial animation system based on data-driven approach. The system can show more than one kind of expressions on the face simultaneously when he/she is talking. We used a combinational approach so that we were able to synthesize different expressions on the different parts of the face coupled with lips synchronization for synthesized speech. The synthesized facial animation can be a combination of the expressions from different faces in databases. We also presented an algorithm to solve the seam problem on the joint area and make the boundary seamless.

This system can be used for various applications. For movie makers, this method may be used as a prototyping tool to producing some expressions with special effect. For game developers, the method can be used to produce some specific virtual roles. By showing some complex expressions on these roles' faces, the player may be deeply impressed.

Acknowledgment

The research was funded by Agency for Science, technology and Research (A*STAR), Singapore. Special thanks go to Hong Thai Nguyen for face model preparation and programming.

References

[1] K. Kahler, J. Haber, and H. P. Seidel, "Geometrybased muscle modeling for facial animation," *In Proceedings of Graphics Interface'2001*, pp. 37–46, 2001.

[2] P. Ekman and W. V. Friesen, "Facial action coding system: a technique for the measurement of facial movement," Consulting Psychologists Press, Palo Alto, CA., 1978.

[3] P. Kalra, A. Mangili, N. M. Thalmann, and D. Thalmann, "Simulation of facial muscle actions based on rational free from deformations," *Eurographics*, vol. 11, no. 3, pp. 59–69, 1992.

[4] K. Kahler, J. Haber, and H.P. Seidel, "Physically-based facial modeling, analysis, and animation," *Proc. Graphics Interface Conf.*, 2001.

[5] D. Terzopoulos and K. Waters, "Physically-based facial modeling, analysis, and animation," *J. Visualization and Computer Animation*, vol. 1, no. 4, pp. 73–80, 1990.

[6] Y. Tang, M. Xu, and Z. Cai, "Research on facial expression animation based on 2d mesh morphing driven by pseudo muscle model," *International Conference on Educational and Information Technology (ICEIT), 2010*, vol. 2, pp. 403–407, 2010.

[7] Y. Cao, W. C. Tien, P. Faloutsos, and F. Pighin, "Expressive speech-driven facial animation," *ACM Transactions on Graphics*, vol. 24, no. 4, pp. 1283 – 1302, Oct 2005.

[8] Z. Deng, U. Neumann, J. P. Lewis, T. Y. Kim, M. Bulut, and S. Narayanan, "Expressive facial animation synthesis by learning speech co-articulations and expression spaces," *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, no. 6, pp. 1523–1534, 2006.

[9] B. Brand, "Voice puppetry," *In Proceedings of ACM SIGGRAPH Conference*, pp. 21–28, 1999.

[10] T. Ezzat, G. Geiger, and T. Poggio, "Trainable videorealistic speech animation,," *ACM Transactions on Graphics*, vol. 21, no. 3, pp. 388–398, 2002.

[11] S. Kshirsagar and N. M. Thalmann, "Visyllable based speech animation," *Computer Graphics Forum (Proc. Eurographics Conf.)*, vol. 22, no. 3, 2003.

[12] E. Chuang and C. Bregler, "Moodswings: Expressive speech animation," *ACM Transactions on Graphics*, vol. 24, no. 2, pp. 331C347, 2005.

[13] E. Chuang, H. Deshpande, and C. Bregler, "Facial expression space learning," *In Proceedings of Pacific Graphics2002*, pp. 68–76, 2002.

[14] Z. Deng and U. Neumann, "Expressive speech animation synthesis with phoneme-level controls," *Computer Graphics Forum*, vol. 27, no. 8, pp. 2096–2113, 2008.

[15] Z. Deng, J.P. Lewis, and U. Neumann, "Automated eye motion synthesis using texture synthesis," *IEEE Computer Graphics and Applications*, vol. 25, no. 2, pp. 24–30, 2005.

[16] G. Kalberer and L.V. Gool, "Face animation based on observed 3d speech dynamics," *Proc. IEEE Computer Animation Conf.*, pp. 20–27, 2001.

[17] P. Cosi, C.E. Magno, G. Perlin, and C. Zmarich, "Labial coarticulation modeling for realistic facial animation," *Proc. Intl Conf. Multimodal Interfaces*, pp. 505–510, 2002.

[18] S. A. King and R. E. Parent, "Creating speech-synchronized animation," *IEEE Trans. Visualization and Computer Graphics*, vol. 11, no. 2, pp. 341–352, 2005.

[19] C. S. Chan and F. S. Tsai, “Computer Animation of Facial Emotions,” in *International Conference on Cyberworlds (CW), 2010*, 2010, pp. 425–429.

[20] J. Ma, R. Cole, B. Pellom, W. Ward, and B. Wise, “Accurate automatic visible speech synthesis of arbitrary 3d model based on concatenation of diviseme motion capture data,” *Computer Animation and Virtual Worlds*, vol. 15, pp. 1–17, 2004.

[21] Y. Cao, P. Faloutsos, E. Kohler, and F. Pighin, “Real-time speech motion synthesis from recorded motions,” *In Proceedings of Symposium on Computer Animation*, pp. 345–353, 2004.

[22] E. Cosatto and H. P. Graf, “Audio-visual unit selection for the synthesis of photo-realistic talking-heads,” *In Proceedings of ICME*, p. 619C622, 2000.

[23] J. Ma, R. Cole, B. Pellom, W. Ward, and B. Wise, “Accurate automatic visible speech synthesis of arbitrary 3d model based on concatenation of diviseme motion capture data,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 15, no. 5, pp. 485C500, 2005.

[24] A. Niswar, E. P. Ong, H. T. Nguyen, and Z. Huang, “Real-time 3D Talking Head from a Synthetic Viseme Dataset,” in *Proceedings of Virtual-Reality Continuum and its Applications in Industry (VRCAI) 2009*. ACM, 2009, pp. 29–33.

[25] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C: The Art of Scientific Computing*, Cambridge University Press, 1992.

[26] L. Liang, C. Liu, Y.Q. Xu, B. Guo, and H.Y. Shum, “Real-time texture synthesis by patch-based sampling,” *ACM Transactions on Graphics*, vol. 20, no. 3, pp. 127–150, 2001.

[27] H. T. Nguyen, E. P. Ong, A. Niswar, Z. Huang, and S. Rahardja, “Automatic and real-time 3d face synthesis,” in *Proceedings of the 8th International Conference on Virtual Reality Continuum and its Applications in Industry 2009*. ACM, 2009, pp. 103–106.