WeCard: A Multimodal Solution for Making Personalized Electronic Greeting Cards

Huijie Lin[1,2], Jia Jia[1,2], Hanyu Liao[3], Lianhong Cai[1,2]
[1]Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
[2] Tsinghua National Laboratory for Information Science and Technology（TNList）
[3]Academy of Art and Design, Tsinghua University, Beijing 100084, China
{linhuijie@gmail.com,jjia@tsinghua.edu.cn,liaoliao1992@gmail.com,clh-dcs@tsinghua.edu.cn}

ABSTRACT

In this demo, we build a practical system, WeCard, to generate personalized multimodal electronic greeting cards based on parametric emotional talking avatar synthesis technologies. Given user-input greeting text and facial image, WeCard intelligently and automatically generate the personalized speech with expressive lip-motion synchronized facial animation. Besides the parametric talking avatar synthesis, WeCard incorporates two key technologies: 1) automatical face mesh generation algorithm based on MPEG-4 FAPs (Facial Animation Parameters) extracted by the face alignment algorithm; 2) emotional audio-visual speech synchronization algorithm based on DBN. More specifically, WeCard merges the users' preferred electronic card scene with emotional talking avatar animation, turning the final content into flash or video file that can be easily shared with friends. By this way, WeCard can help you make your multimodal greetings to be more attractive, beautiful, and sincere.

Categories and Subject Descriptors

H.5.1 [**Information Interfaces and Presentation**]: Multimedia Information Systems—Evaluation/methodology

Keywords: Personalized, Talking Avatar, Emotion, Multimodal

1. INTRODUCTION

Talking avatar is an effective way to provide natural human-like communication via computers. During the past years, much effort has been devoted to enhance the performance of speech synthesis and facial animation. However, most existing talking avatars are based on data-driven methods, which is difficult to adopt personalized face model to achieve users' specific facial animation. On the other hand, most of the existing talking avatars focus on the synchronization of speech and lip motion, while the emotion synchronization of speech and facial animation is the same important.

In this demo, we present a practical application based on parametric emotional talking avatar synthesis technology [1,2] to automatically create personalized greeting e-cards with multimodal content, named WeCard. Based on the proposed efficient face mesh generation algorithm and real-time emotional audio-visual synchronization algorithm [1], WeCard needs only several easy steps to create the e-card. A user just need to type the greeting text and upload a selected facial image (even the cartoon image), the WeCard will intelligently and automatically generate an e-card in form of video or flash presenting the users' specified speech style with synchronized facial expressions and lip motions.

2.SYSTEM INTERFACE AND WORKFLOW

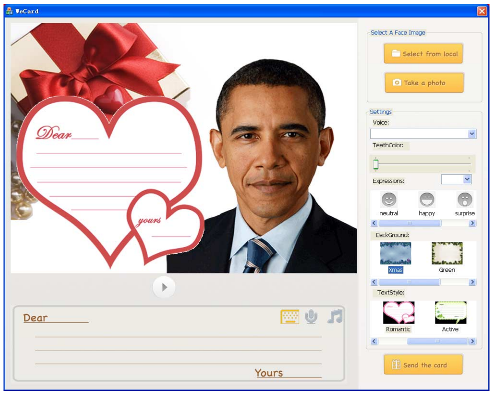

Figure 1. System Interface of the WeCard.

As shown in Figure 1, to generate the electronic greeting card, the following interaction steps are needed:1) User first select a facial image from the local disk or take a photo from the camera, and an **automatical face mesh generation algorithm** will be applied to generate the MPEG-4 standard compliant face mesh; 2) Input the greeting text or record a greeting audio or song, WeCard will first synthesize the text to speech with a Multilingual TTS (Chinese/English) and generate time-stamped phoneme series, or recognize the phoneme series directly from the input speech. Then the phoneme series are turned into FAPs (Facial Animation Parameters) using a **bilingual FAPs synthesizer** [1]; 3) Select the preferred speech style, such as cool man, young lady, boy, girl, robot, etc. WeCard will modify the pitch and spectrum simultaneously to achieve the specified speech style using the **LP-PSOLA based voice conversion method** [1**]**; 4) Select the expression of the facial animation, including happy, surprise, funny, etc. the facial expression will be generated by the **PAD** (Pleasure-Arousal-Dominance psychological model) **- FAPs mapping model** [1,2]. The **PAD** values used to generate **FAPs** are presetted according to the selected emotion category. 5) WeCard will automatically analyze the acoustic features of the input speech and do the **emotional audio-visual speech synchronization** to generate emotional-lip-synchronized FAPs [1]; 6) Choose an appropriate scene from the card background database; Finally, click the "Preview" button to check the generated e-card in the preview window. After that, the final e-card can be exported and shared with friends.

3. ALGORITHMS

WeCard is established based on parametric talking avatar synthesis [1,2,5]. Besides, we proposed an **automatical face mesh**

MM'13, October 21–25, 2013, Barcelona, Spain.
ACM 978-1-4503-2404-5/13/10.

generation algorithm to deal with arbitrary facial images, and **emotional audio-visual speech synchronization algorithm** to synchronize the speech, viseme and facial expression.

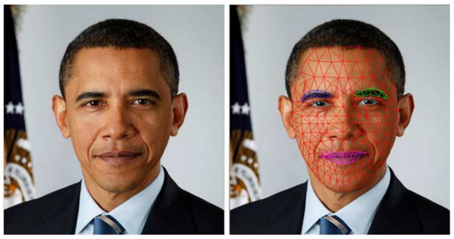

Figure 2. Face mesh auto-generated from photograph.

Automatical face mesh generation algorithm. To transform a static face image to dynamic facial animation, the input facial image should first be parameterized to Facial Definition Parameters defined by MPEG-4 standard. First, a standard face mesh is predefined to model the general face as shown in Figure 3. Then we extract 88 feature points from the input face image using the face alignment algorithm in [3], which can precisely describe the contour and the main parts of the face, including the eyebrows, eyes, nose and mouth. A quadratic interpolation function is applied to generate the curves of the contour and main parts of the face according to the extracted feature points. After that, a precise and topological isomorphic mapping is established between the standard face mesh and the generated feature curves. Then the standard face mesh is deformed to fit the input face according to the mapping information. Thus the final face model satisfying the MPEG-4 standard is established.

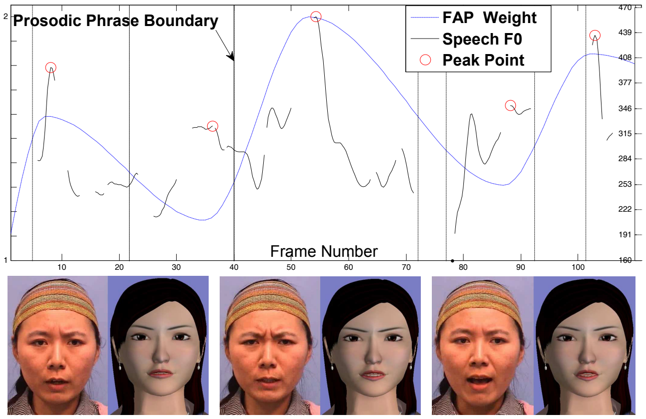

Figure 4. FAPs synchronized by acoustic features (Anger, P: -0.69, A: 0.51, D:0.11). Compare the synthetic avatar with human video [1].

Emotional audio-visual speech synchronization. We proposed a DBN-based audio-visual correlative model (AVCM)[1], where the loose timing synchronicity between audio and video streams is restricted by word boundaries. For the audio part, the emotional speech is converted by LP-PSOLA algorithm from the neutral speech synthesized by TTS. For the video part, since the mouth movement plays an important role in both speaking and facial expression, we take a linear weighted function to merge the animation parameter of viseme and facial expression in mouth region [1]. For audio-visual synchronization, the inputs of DBN based AVCM are: 1) acoustic features extracted from emotional speech, and 2) the FAPs generated by merging the facial expression with viseme. The previously-trained model is applied to calculate the probability score, which is used as a measure of synchronization error. The downhill simplex method is used to adjust the FAPs according to the change of acoustic features until we got the smallest synchronization error. Thus, we will get the emotional and lip motion synchronized audio-visual speech.

4. EXPERIMENTAL RESULTS

To evaluate the effectiveness of our proposed algorithm and emotional talking avatar synthesis system, objective and subjective experiments are conducted.

- **Automatical face mesh generation algorithm**. We invited 10 test users to take their front face photos and annotated the face meshes to compare the key parts with that automatically generated by our system. Results show that the mean average accuracy of our automatically generated face mesh is around 94.8%.
- **Emotional audio-visual speech synchronization**. To test the emotional audio-visual speech synthesis, we first randomly select 50 audio-visual utterances from a labeled dataset containing 132 emotional audio-visual speeches [1], each of which is annotated with PAD values. Using the PAD values and the corresponding texts as inputs, we synthesize 50 audio-visual utterances on a human facial image. We invite ten participants to compare the synthetic audio-visual speeches to the recorded speeches with the same PAD values and texts. We use the average MOS values to describe the similarity between the emotional audio-visual speeches and the corresponding videos (highest:5; lowest:1). The MOS average score is 3.4, which indicates our system can synthesize a natural and expressive emotional audio-visual speech.

5. CONCLUSIONS

In this demo, we build a practical system, WeCard, to generate personalized multimodal electronic greeting cards. Based on parametric talking avatar synthesis [1,2], WeCard incorporates two other key technologies: 1) automatical face mesh generation algorithm based on MPEG-4 FAPs extracted by the face alignment algorithm; 2) emotional audio-visual speech synchronization algorithm based on DBN. Furthermore, WeCard merges the users' preferred electronic card scene with the emotional talking avatar animation, turning the final content into flash or video file. The personalized electronic greeting cards generated by WeCard can be easily shared with friends, which makes your greetings to be more attractive, beautiful, and sincere.

6. ACKNOWLEDGEMENTS

This work is supported by the 973 Program of China (2011CB302201), and is partially supported by the 973 Program (2012CB316401) and 863 Program (2012AA011602). The work is also funded by Microsoft Research Asia-Tsinghua University Joint Laboratory.

7. REFERENCES

[1] Jia Jia, Shen Zhang, Fanbo Meng, et al. Emotional Audio-Visual Speech Synthesis based on PAD. IEEE Transaction on Audio, Speech, and Language Processing, Vol.19 No.3 pp570-582, 201.

[2] Jia Jia, Xiaohui Wang, Zhiyong Wu, et al. Modeling the Correlation between Modality Semantics and Facial Expressions. APSIPA, Hollywood, California, Dec 3-6, 2012

[3] Li Zhang, Haizhou Ai, et al. Robust Face Alignment Based on Local Texture Classifiers. IEEE ICIP 2005, Genoa, Italy, 2005.

[4] Roberto Valenti, Alejandro Jaimes, Nicu Sebe. Sonify Your Face: Facial Expressions for Sound Generation.The International Conference on Multimedia, pp1363-1372, 2010.

[5] Jia Jia, Zhiyong Wu, Shen Zhang, et al. Head and facial gestures synthesis using PAD model for an expressive talking avatar. Multimedia Tools and Applications, Springer, DOI: 10.1007/S11042-013-1604-8.