Text Driven 3D Photo-Realistic Talking Head

Lijuan Wang[1], *Wei Han*[1,2], *Frank K. Soong*[1], *Qiang Huo*[1]

[1] Microsoft Research Asia, Beijing, China
[2]Department of Computer Science, Shanghai Jiao Tong University, China
lijuanw@microsoft.com, weihan_cs@sjtu.edu.cn,
frankkps@microsoft.com, qianghuo@microsoft.com

Abstract

We propose a new 3D photo-realistic talking head with a personalized, photo realistic appearance. Different head motions and facial expressions can be freely controlled and rendered. It extends our prior, high-quality, 2D photo-realistic talking head to 3D. Around 20-minutes of audio-visual 2D video are first recorded with read prompted sentences spoken by a speaker. We use a 2D-to-3D reconstruction algorithm to automatically adapt a general 3D head mesh model to the individual. In training, super feature vectors consisting of 3D geometry, texture and speech are formed to train a statistical, multi-streamed, Hidden Markov Model (HMM). The HMM is then used to synthesize both the trajectories of geometry animation and dynamic texture. The 3D talking head animation can be controlled by the rendered geometric trajectory while the facial expressions and articulator movements are rendered with the dynamic 2D image sequences. Head motions and facial expression can also be separately controlled by manipulating corresponding parameters. The new 3D talking head has many useful applications such as voice-agent, tele-presence, gaming, social networking, etc.

Index Terms: audio/visual synthesis, 3D, photo-realistic, talking head

1. Introduction

Avatars can be roughly divided into two categories, depending upon how avatars interact with the outside world: the first one plays in human-to-human communications, like tele-presence; the other one acts as an intelligent agent in human-computer interactions. Some desirable features of the next generation avatar are: it should be a 3D Avatar to be integrated easily into a versatile 3D virtual world; it should be photo-realistic; it can be customized to any user; last but not least, an avatar should be automatically created with a small amount of recorded data. In summary, the next generation Avatar should be 3D, photo-realistic, personalized or customized, and easy to create with little bootstrapping data. They are the ultimate goal of this 3D photo-realistic project.

A traditional 3D avatar requires a highly accurate geometric model to render soft tissues like lips, tongue, facial expressions, wrinkles, etc. It is both computationally intensive and mathematically challenging to make or run such a model. Moreover, any unnatural deformation will make the resultant output fall into the “uncanny valley” of human rejection. That is, it will be rejected as un-natural. On the other hand, in a 2D talking head rendered by concatenating short video segments [1-4], it is challenging to change the head pose freely or to render different facial expressions. Additionally, it is very hard to blend a 2D talking head in a 3D view seamlessly.

Our new 3D photo-realistic talking head takes the advantages of both 2D and 3D avatars. It renders a 3D head by mapping or wrapping 2D video images around a simple, 3D mesh model. The 2D video can capture the natural movement of soft tissues (e.g. lips and tongue) and it helps the new talking head to bypass the problem caused by occluded articulators (e.g. tongue and teeth). The simple 3D model allows us to render any rigid body motion of a head. Therefore, this new 3D photo-realistic talking head combines the advantage of a 2D video and 3D mesh model and it overcomes the “uncanny valley” problem of unnatural rendering. The result shows that the 3D photo-realistic talking head is natural and acceptable to human eyes.

2. Synthesizing 2D Photo-Realistic Lips Movement

Fig. 1 shows the flow of our 2D photo-realistic talking head synthesis system, which consists of two, training and synthesis, stages [5-8].

In the training stage, audio/visual footage of a speaker is used to train the statistical audio-visual Hidden Markov Model (AV-HMM). The input of the HMM contains both the acoustic features and the visual features. The acoustic features consist of Mel-frequency cepstral coefficients (MFCCs), their delta and delta-delta coefficients. The visual features include the PCA coefficients and their dynamic features., The contextual dependent HMM is used to capture the variations caused by different contextual features. Also, the decision tree-based clustering technique is applied to cluster the acoustic and visual states into tied states respectively to improve the robustness of the HMM. The audio/visual HMM modeling is firstly trained with the traditional maximum likelihood (ML) estimation, then jointly refined by using a probabilistic descent algorithm to optimize the model parameters under the Minimum Generation Error (MGE) criterion. The MGE training can explicitly optimize the quality of generated visual speech trajectory.

In the synthesis stage, the input phoneme labels and alignments are firstly converted to a context-dependent label sequence. Meanwhile, the decision trees generated in the training stage are used to choose the appropriate clustered HMM states for each label. Then parameter generation algorithm is used to generate the visual parameter trajectory in maximum likelihood sense. The HMM predicted trajectory is used as guidance for selecting a succinct mouth sample sequence from the image library. Finally the mouth image sequence is stitched onto the background head video. The synthesized 2D face video will be projected to a 3D head mesh model for rendering the 3D photo-realistic talking head.

3. 3D Photo-Realistic Talking Head

Traditional 3D avatar requires a highly accurate geometric model to render realistic facial animations. However, estimating geometric structure from uncalibrated images accurately enough for high quality 3D rendering is difficult, especially for human face which everyone is familiar with.

Even when given an accurate 3D face model, it is still difficult to deform the model properly for different facial motions, especially for the non-rigid (soft) tissues on the face like lips, tongue, eyes, wrinkles, etc. By using the 2D-to-3D reconstruction algorithm in [9-11], a simple and smooth face mesh model is estimated from a single frontal face image or a short 2D video, as shown in Fig. 2. Instead of using a precise model of the geometry mesh deformation, local facial motion is obtained by overlaying a dynamic, time varying texture on the simple 3D head mesh. Unlike traditional texture mapping which generates a single texture for a surface, multiple textures are used in rendering by using dynamic texture mapping. Then, we synthesize a 2D video of a real person and obtain a sequence of images of the mouth movements. We project the images of the mouth on the 3D head, which is a simple and smooth model. As the mouth opens and closes in the 2D video, its projection on the 3D head also opens and closes. Also, the projection can be observed in different direction. Therefore, it can bypass the difficulties in rendering soft tissues like lips, tongue, eyes, wrinkles, and make the 3D talking head look photo-realistic.

With the versatile 3D geometry model, the head pose, illumination, and facial expressions of the 3D talking head can be freely controlled. In particular, head movement can be controlled by rotating and translating the head mesh model by viewing it as a rigid object. Different illumination can be realized by changing the lighting in 3D rendering. Various facial expressions like happy or sad can be controlled by deforming the 3D mesh model.

Fig. 3 shows the snapshot of the 3D photo-realistic talking head in different head poses and facial expressions. A demo video is also enclosed to demonstrate the results of the 3D photo-realistic talking head (http://research.microsoft.com/en-us/projects/photo-real_talking_head/3d_intro_2.avi).

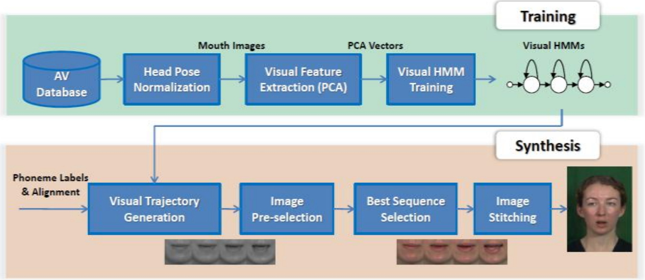

Figure 1: 2D Photo-Realistic lips movement synthesis.

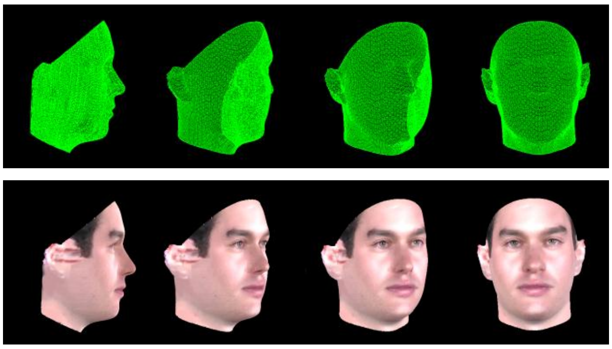

Figure 2: Auto-reconstructed 3D face model (w/o and w/ texture) in different poses.

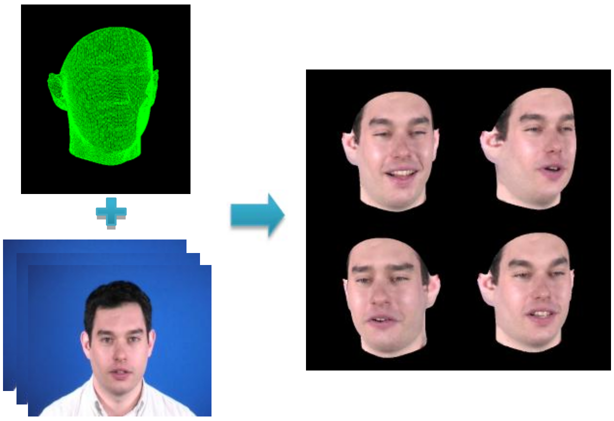

Figure 3: 3D Photo-Realistic Talking head.

4. Conclusion

We create a digital, 3D photo-realistic, personalized Avatar to let you "see yourself" in a virtual world experience. Our method combines the advantages of both 2D and 3D avatars. Firstly, it is 3D. The 3D head can be rotated easily. Secondly, it is simple. There is no complex modeling of lips and muscles required. Thirdly, it is photo-realistic. The 2D images capture the natural movement of soft tissues. The projection on the 3D head is also natural. So, it overcomes the "uncanny valley" problem. Fourthly, our avatar can be customized to any user by using the 2D video of the user. This offers a way to create you personalized 3D photo-realistic talking head with a short 2D video clip.

5. References

[1] E. Cosatto and H. P. Graf, "Photo-Realistic Talking Heads From Image Samples," *IEEE Trans. Multimedia*, vol. 2, no. 3, pp. 152-163, 2000.

[2] T. Ezzat, G. Geiger, and T. Poggio, "Trainable Video Realistic Speech Animation," in *Proc. ACM SIGGRAPH2002*, San Antonio, Texas, 2002, pp. 388-398.

[3] B. J. Theobald, S. Fagel, G. Bailly, and F. Elisei, "LIPS 2008: Visual Speech Synthesis Challenge," in *Proc. INTERSPEECH* 2008, pp. 2310-2313.

[4] K. Liu, J. Ostermann, "Realistic Facial Animation System for Interactive Services," in *Proc. INTERSPEECH 2008*, Brisbane, Australia, Sept. 2008, pp.2330-2333.

[5] L. Wang, Y. Wu, X. Zhuang, and F. Soong, "Synthesizing Visual Speech Trajectory with Minimum Generation Error," in *Proc. ICASSP 2011,* Prague, Czech Republic, May 2011.

[6] L. Wang, W. Han, X. Qian, and F. Soong, "Synthesizing Photo-Real Talking Head via Trajectory-Guided Sample Selection," in *Proc. INTERSPEECH 2010*, Chiba, Japan, Sept. 2010, pp.446-449.

[7] X. Zhuang, L. Wang, F. Soong, and M. Hasegawa-Johnson, "A Minimum Converted Trajectory Error (MCTE) Approach to High Quality Speech-to-Lips Conversion," in *Proc. INTERSPEECH 2010*, Chiba, Japan, Sept. 2010, pp.1736-1739.

[8] K. Wu, L. Wang, F. Soong, Y. Yam, "A Sparse And Low-Rank Approach To Efficient Face Alignment For Photo-Real Talking Head Synthesis," in *Proc. ICASSP 2011,* Prague, Czech Republic, May 2011.

[9] Y. Hu, D. Jiang, S. Yan, L. Zhang, H. Zhang, "Automatic 3D Reconstruction for Face Recognition," in *Proc. of the Sixth IEEE international Conference on Automatic Face and Gesture Recognition (FGR'04)*.

[10] L. Xin, Q. Wang, J. Tao, X. Tang, T. Tan, and H. Shum, "Automatic 3D Face Modeling from Video," in *Proc. ICCV 2005*, Oct. 2005, pp.1193-1199.

[11] Z. Liu, Z. Zhang, D. Adler, E. Hanson, M. Cohen, "A Robust and Fast Face Modeling System," in *Proc. IEEE Pacific Rim Conference on Multimedia 2001*, pp.269-276.