Speech Driven Photo-Realistic Face Animation with Mouth and Jaw Dynamics

Ying He[1,2], Yong Zhao[1,2], Dongmei Jiang[1,2]
VUB-NPU Joint Research Group on AVSP
1. Northwestern Polytechnical University
2. Shaanxi Provincial Key Laboratory of Speech and Image Information Processing, Xi'an, P. R. China
he_ying@mail.com, jiangdm@nwpu.edu.cn,

Hichem Sahli[1,2*]
1 VUB-NPU Joint Research Group on AVSP
Electronics & Informatics Dept - ETRO
Vrije Universiteit Brussel (VUB)
2 Interuniversity Microelectronics Centre – IMEC
Brussels, Belgium
hsahli@vub.ac.be

Abstract— **This paper proposes a system that transforms speech waveform to photo-realistic speech-synchronized talking face animations. We expand the multi-modal diviseme unit selection based mouth animation system of [8] to a full photo realistic facial animation system based on (i) modeling of the non-rigid deformations of the mouth and jaw via a general regression neural network, (ii) multi-resolution image blending approach for fusing the synthesized mouth image to the full face image, and (iii) synthesizing natural head poses or deflections using a modified version of the generalized procrustes analysis for face image alignment. The paper describes the main principles of the proposed method and illustrates its results on a set of testing speech sequences, together with qualitative and quantitative comparisons with results from the approach of the recognized system Video Rewrite. Experimental results show that the proposed method obtains realistic facial animations with very natural mouth and jaw movements coincident with the input speech.**

Keywords- Face image alignment, Diviseme unit selection, General regression neural network, Multi-resolution image blending.

I. INTRODUCTION

Computer simulation of human faces capable of reflecting mouth movements has been a flourishing research area during the last decade. A large number of facial models and several talking face animation systems have been built. According to the underlying face model, these systems can be categorized into 3D-model-based animation or image-based rendering systems. While 3D graphical models [1, 2] provide parametric control but lack of video realism, image-based approaches [6-8] have the potential of achieving very high levels of video realism.

Within the image based photo realistic facial animation systems, some approaches adopt the parameterization methods for the mouth (facial) images, which are capable of synthesizing the previously unseen mouth configurations from a small set of mouth image prototypes. Among them, machine learning strategy considers mouth synching as an audio-to-visual conversion problem [3-5]. Clear and natural mouth images can be constructed from the visual features learned by an audio visual speech model. However, since the visual feature learning is based on the Maximum Likelihood Estimation (MLE) giving the global optimal solution, the accuracy and details of the synthesized mouth shapes are not promising. On the other hand, [6] proposes a variant of the multidimensional morphable model (MMM) of the mouth images. A post-processing stage composites the mouth movement onto a background sequence containing natural eye and head movements. However, since the generated mouth mask contains the region from the lower eyelid to the neck, it has the difficulty of re-compositing the synthesized mouth sequences into background sequences which involve large changes in head pose, or changes in lighting conditions.

Most of the video realistic talking face systems adopt the unit selection and concatenation strategy. Video Rewrite [7] is one of the recognized examples, in which a new audiovisual sentence is constructed by concatenating together the appropriate triphone sequences from the database. In stitching the triphone videos into the background face image sequence, a replacement mask is used to specify which portions of the final video come from the triphone images and which come from the background video. Very realistic facial animations are obtained, but since the mask is relatively large to include the mouth and chin, if the lighting conditions on the synthesized mouth and the upper face are different, the overall naturalness of the animation will be influenced.

In [8], a speech driven mouth synthesis method based on multi-modal diviseme unit selection has been proposed. An audio visual diviseme database consisting of the acoustic feature and intensity sequences, as well the visual feature sequences of the diviseme instances, is first build. Then a diviseme instance selection algorithm is introduced to select the optimal representative diviseme instances for the viseme pairs in the input speech. The image sequences of the final selected diviseme instances are time warped and blended to construct the new images of the mouth animation. Experimental results showed that the multimodal diviseme instance selection scheme of [8] outperforms the triphone unit selection algorithm in Video Rewrite. Clear, accurate, smooth mouth animations can be obtained matching well the pronunciation and intensity changes of the incoming speech.

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 EU FP7 project ALIZ-E (grant 248116), and the VUB-HOA CaDE project.

In this paper, we expand the multi-modal diviseme unit selection based mouth animation system of [8] to a full facial animation system. The novelties of this work are: 1) to reduce the influence of head pose and face deflection, we align the face images as well their feature points to a reference face by a generalized procrustes analysis (GPA) method; 2) in order to 'align' jaw shapes to be coincident with the mouth movements in the synthesized facial animation, we identify the mapping from mouths deformations to jaws, and hence the non-rigid deformations of mouth and jaws, using the Generalized Regression Neural Networks (GRNN). The mouth feature points, obtained from the mouth image constructed by the diviseme unit selection method, are input to the GRNN to obtain the jaw points, which are then used in the morphing of a background face image. 3) Finally the synthesized mouth images and the morphed face images are fused via multiresolution image blending for natural appearance in synchrony with the input speech. The mask of the fusion process is generated by determining the convex hull of mouth/jaw region using the upper lip feature points and the three lowest jaw feature points. This relatively small fusion mask makes the synthesized face animation more robust to the lighting condition changes in the database. Unlike the rigid background face in [6], the background face as well the synthesized face of the proposed approach could move with different head poses or deflections. Objective and subjective evaluations show that the proposed method outperforms Video Rewrite in both jaw movements and mouth (jaw) naturalness.

The rest of the paper is organized as follows: Section II describes the alignment of the face images in the database, as well training of the non-rigid deformations of the mouth and jaw positions using GRNN. Section III discusses the morphing of the background face images, and Section IV the final fusion of the synthesized mouth image with the morphed background face image. Experimental results are given in Section V. In section VI, we draw conclusions.

II. Face Image Alignment and Modeling of the Non-rigid Deformations of the Mouth and Jaw

The constrained Bayesian tangent shape model (CSM) of [9] has been used for the detection and tracking of a shape model defined by 83 facial feature points, over a facial image sequence. Fig.1 shows the result of one image frame, together with the locations of the feature points.

To take into account head movements and appearance for the training of mouth-jaw non-rigid relationship, the Generalized Procrustes Analysis (GPA) method [10] has been applied for aligning all the face images in the database. Since the procrustes transformation, being based on similarity transformation, is not able to correct the out-of-plane distortion, we modified the GPA algorithm by adopting an affine transformation to align two shapes. The process is as follows: Firstly, we randomly select from the database an initial reference face, and align all the faces to the reference face using an affine transformation. A new reference face is then estimated as the mean of all the aligned faces, and the alignment process is repeated until the difference between the aligned face and the reference face is small. As a side product, a final reference face is also produced. Fig.2 illustrates the results of the modified GPA based alignment approach, which allows correcting the pose and deflection of the face.

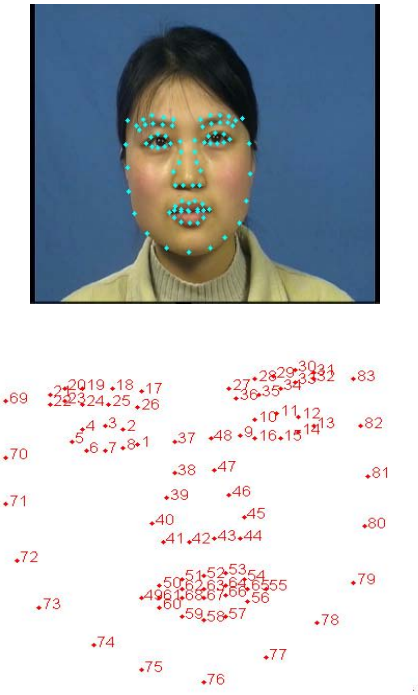


Figure 1. Face detection/tracking results and locations of the 83 facial points

Figure 2. Face image alignment. Left: original face; Right: aligned face

After alignment of the face images, as well their feature points, the general regression neural network (GRNN) is adopted to model the non-rigid deformation of the mouth and jaw feature points. The 20 feature points of the lip contour (points 49-68 in Fig.1) are used as inputs of the GRNN, and the 9 feature points of jaw (points 72-80 in Fig.1) are its outputs. In our experiment, mouth and jaw feature points of 4000 face images, randomly chosen from the audio visual speech database, are used to train the GRNN.

III. Morphing of the Background Face Images

From each aligned face image of the audio visual speech database, a 64x72 mouth region of interest (ROI) is extracted, based on the 12 feature points of the outer lip contour. These mouth images are then used to construct an audio visual diviseme (viseme pair) unit database. For an input audio speech, the multi-modal diviseme unit selection method [8] is adopted

by considering the smoothness of the synthesized mouth movements, as well as the similarity of intensity and pronunciation between the input speech and the diviseme unit. The mouth image sequences of the selected diviseme units are then time warped and concatenated. As a result, a new mouth image sequence, together with the feature points on the lip contours, is synthesized. The synthesized sequence maps well the pronunciation and intensity of the input speech. For detailed information, please refer to the work in [8].

For the final facial animation, the synthesized mouth image sequence is blended with a background face image sequence extracted from the database. As the jaw shapes in the background face images are probably not coincident with the synthesized mouth shapes (which will influence the naturalness of the facial animation), we morph background face images before the blending of the mouth and face images. This would allow having the jaw shapes mapping the synthesized mouth shapes. The process is done as follows:

The GRNN is used to map the non-rigid relationship between mouth-jaw. For each frame, the 20 feature points of lip contour of the synthesized mouth image are inputs to the GRNN where its output are the 9 learned feature points of jaw, $\left[(x'_{72}, y'_{72}),....(x'_{80}, y'_{80})\right]$.

To make smooth the transition between the learned jaw points and the other feature points of the background face, we set the feature points of jaw as

$$\begin{aligned} \tilde{x}_k &= \alpha_k x'_k + (1-\alpha_k)x_k, \\ \tilde{y}_k &= \alpha_k y'_k + (1-\alpha_k)y_k \end{aligned} \quad , \quad k = 72,...,80 \qquad (1)$$

where (x_k, y_k) are the coordinates of the jaw feature point on the background face. α_k are set empirically giving the highest weight to the lowest jaw point 76, and weak weights to the points near ear, so that the synthesized lower jaw shape follows as closely as possible the learned feature points, and the upper part transits smoothly with the other facial feature points.

Once the new feature point vector of the face

$$\tilde{X} = \left[(x_1, y_1),...,(\tilde{x}_{72}, \tilde{y}_{72}),....(\tilde{x}_{80}, \tilde{y}_{80}),...,(x_{83}, y_{83})\right]$$

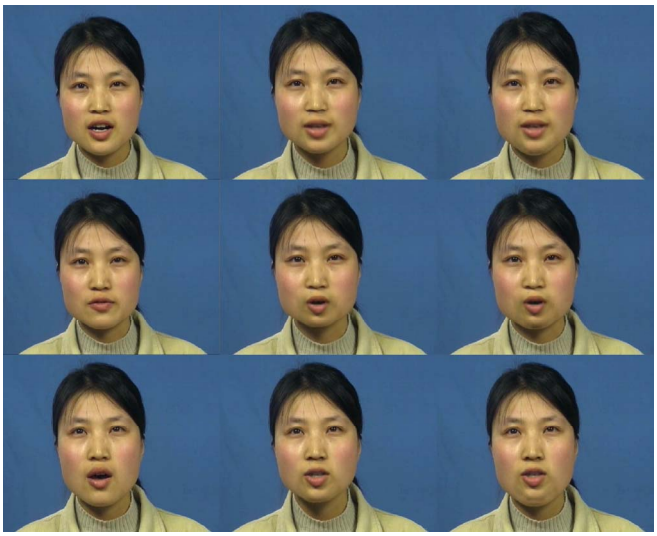
is obtained, the background face image with the feature points

$$X = \left[(x_1, y_1),...,(x_{72}, y_{72}),....(x_{80}, y_{80}),...,(x_{83}, y_{83})\right]$$

is morphed by the piecewise linear mapping algorithm [11]. This algorithm has been selected due to its inherent locality property.

IV. Blending of the Mouth Image with the Background Face Image

The final operation of the proposed photo-realistic synthesis is blending the synthesized mouth image to the morphed background face image. This is made using a multi-resolution Pyramid blending approach of [12], to produce a coherent composite image from a combination of the component images. The mouth mask required for this process is generated by computing the convex hull of the feature points 49-55 of the upper lip, as well as the feature points 75-77 of the jaw. Finally, an inverse affine transformation is applied to the fused image to keep the video background (blue in our case) homogenous. This is indeed required because the mouth synthesis and the background face morphing are applied on the aligned face images. Fig.3 shows some frames of the synthesized face images. As it can be seen from Fig.3.b, the jaw positions map better the mouth shapes when using the background face morphing.



(a) (b) (c)

Figure 3. Synthesized face images. Col. (a): background face image; Col.(b): synthesized face with morphing; Col.(c): synthesized face without morphing.

V. Experiments and Results

In our experiment, 100 audio visual speech sentences of a female speaker are used as training set. Face animations are synthesized for 100 audio speech sentences of a male speaker. As baseline comparison with our experiment, we adopt the results from the Video Rewrite method [7], since such approach is also a speech driven photo realistic face animation system, selecting and concatenating the triphone units and fusing the synthesized mouth (jaw) image to the background face image.

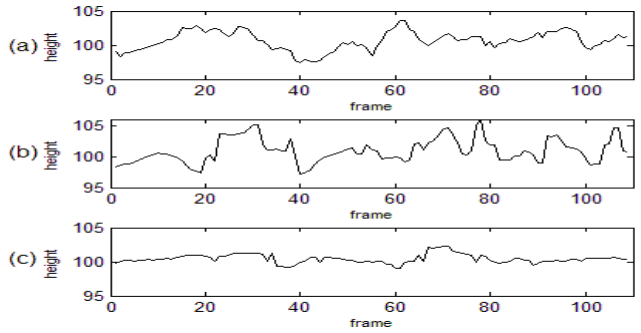
A. Jaw Trajectories

We define the height of jaw as the distance between the nose apex and the point 76 of jaw. Fig.4 shows the jaw height trajectories of two speech sentences obtained from Video Rewrite, and the proposed method, respectively. The ground truth trajectories are obtained from the recorded videos of the driving speech. One can notice that, by learning the jaw feature points using the GRNN, and morphing the background face image, the obtained jaw trajectories track more closely the ground truth trajectories.

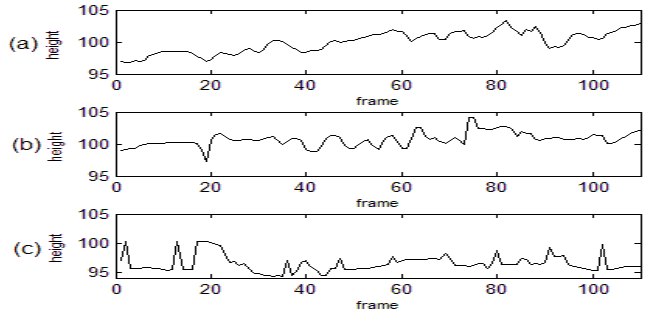
To objectively measure the movement of jaw, we calculate the mean relative distance (MRD) between the ground truth jaw height and the synthesized jaw height as

$$MRD = \frac{1}{M}\sum_{i=1}^{M}\left\{\frac{1}{L_i}\sum_{j=1}^{L_i}\left|[h(j)-mean(i)]-[h_{ori}(j)-mean_{ori}(i)]\right|/h_{ori}(j)\right\} \qquad (2)$$

with M being the number of testing speech sentences (100 in our experiment), L_i the image frame number of the i th speech sentence. $h(j)$ and $h_{ori}(j)$ are the jaw heights of the synthesized face image, and of the ground truth of frame j, respectively. $mean(i)$ and $mean_{ori}(i)$ are the mean jaw heights of the synthesized face image sequence and of the original reference of sentence i. For the proposed approach, the MRD value is 0.4644, and for the Video Rewrite approach, MRD is 0.9190. This shows that by morphing the face images to the learned jaw shapes, the proposed method produces more reasonable jaw movements mapping better the input speech.



(i) Jaw trajectories of speech sentence 1



(ii) Jaw trajectories of speech sentence 2

Figure 4. Jaw trajectories of two speech sentences. (a): ground truth jaw trajectory. (b): jaw trajectory obtained by the proposed approach. (c): jaw trajectory generated by Video Rewrite

B. Subjective Evaluation

TABLE I. STATISTICS ON THE SCORES FOR THE PROPOSED METHOD

Scores	1	2	3	4	5	MOS
Mouth naturalness	15	117	1026	672	170	3.433
Jaw naturalness	0	29	672	694	605	3.938
Overall naturalness	0	30	627	1092	251	3.782

TABLE II. STATISTICS ON THE SCORES FOR VIDEO REWRITE

Scores	1	2	3	4	5	MOS
Mouth naturalness	52	450	1026	428	44	2.981
Jaw naturalness	0	88	871	598	443	3.698
Overall naturalness	0	207	923	745	125	3.394

To subjectively evaluate the synthesized facial animations, 20 relatively inexperienced students are asked to score the 200 synthesized face animations, from the proposed method and from Video Rewrite, respectively. The scores are scaled into five levels: 1(bad), 2(poor), 3(fair), 4(good), 5(excellent), and the Mean of Score (MOS) results are given in Table 1 and Table 2. One can notice that the proposed speech driven facial animation system outperforms the Video Rewrite approach, more natural mouth and jaw movement, mapping better the input speech, can be obtained.

VI. CONCLUSIONS AND FUTURE WORK

This paper presents a novel speech driven face animation synthesis method. To reduce the influence of head pose and face deflection, the face images in the database are aligned to a reference face, then the mapping of the non-rigid deformations of the mouth and jaw are obtained via a GRNN, which are then used to morph the background face image. Finally, the synthesized mouth image is blended with the morphed face image to construct the final face image for the animation. Objective and subjective evaluations show that the proposed method outperforms the Video Rewrite approach in both jaw movements and mouth (jaw) naturalness. In our future work, we would like to integrate our emotion recognition work [13] with the proposed facial animation synthesis method, so that a talking face animation not only shows the mouth and jaw movements coincident with the speech content, but also with facial expressions mapping the emotion in the speech.

REFERENCES

[1] G. Salvi, J. Beskow, S.A. Moubayed, and B. Granstrm, "SynFace - Speech-Driven Facial Animation for Virtual Speech-Reading Support", presented at EURASIP J. Audio, Speech and Music Processing, 2009.

[2] Ning Liu; Ning Fang; Kamata, S.; , "3D reconstruction from a single image for a Chinese talking face," TENCON 2010 - 2010 IEEE Region 10 Conference , vol., no., pp.1613-1616, 21-24 Nov. 2010

[3] K. Choi, Y. Luo, and J. Hwang, "Hidden Markov Model Inversion for Audio-to-Visual Conversion in an MPEG-4 Facial Animation System", presented at VLSI Signal Processing, 2001, pp.51-61.

[4] L. Xie and Z. Liu, "A coupled HMM approach to video-realistic speech animation", presented at Pattern Recognition, 2007, pp.2325-2340.

[5] D. Jiang, I. Ravyse, P. Liu, H. Sahli, and W. Verhelst, "Realistic mouth animation based on an articulatory DBN model with constrained asynchrony", in Proc. ICASSP, 2010, pp.2478-2481.

[6] T. Ezzat, G. Geiger, and T. Poggio, "Trainable videorealistic speech animation", presented at ACM Trans. Graph., 2002, pp.388-398.

[7] C. Bregler, M. Covell, and M. Slaney, "Video Rewrite: driving visual speech with audio", in Proc. SIGGRAPH, 1997, pp.353-360.

[8] D. Jiang, I. Ravyse, H. Sahli, and W. Verhelst, "Speech driven realistic mouth animation based on multi-modal unit selection," Journal on Multimodal User Interfaces, vol. 2, pp. 157-169, 2008.

[9] Y. Hou, H. Sahli, R. Ilse, Y. Zhang, and R. Zhao, "Robust Shape-Based Head Tracking," Advanced Concepts for Intelligent Vision Systems, vol. 4678, pp. 340-351, 2007.

[10] J. Gower, "Generalized procrustes analysis," Psychometrika, vol. 40, pp. 33-51, 1975.

[11] A. Goshtasby, "Piecewise linear mapping functions for image registration," Pattern Recognition, vol. 19, pp. 459-466, 1986.

[12] P. J. Burt and E. H. Adelson, "A multiresolution spline with application to image mosaics," ACM Trans. Graph., vol. 2, pp. 217-236, 1983.

[13] D. Jiang, Y. Cui, X. Zhang, P. Fan, I. Ganzalez, and H. Sahli, "Audio visual emotion recognition based on triple-stream dynamic bayesian network models," in Proceedings of the 4th international conference on Affective computing and intelligent interaction, Memphis, TN, 2011, vol. 1, pp. 609-618.