Video Reshuffling: Automatic Video Dubbing without Prior Knowledge

Shoichi Furukawa[†1] Takuya Kato[†1] Pavel Savkin[†1] Shigeo Morishima[†2]

Waseda University[†1] Waseda Research Institute for Science and Engineering[†2]

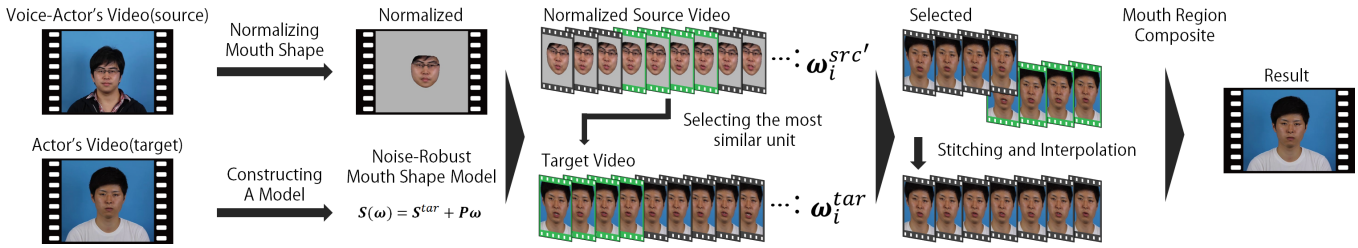

Figure 1: *Outline*

Keywords: Computer Graphics, Face Synthesis, Face Analysis

Concepts: •**Computing methodologies** → *Computer graphics;*

1 Introduction

Numerous video have been translated using "dubbing," spurred by the recent growth of video market. However, it is very difficult to achieve the visual-audio synchronization. That is to say in general a new audio does not synchronize with actor's mouth motion. This discrepancy can disturb comprehension of video contents. Therefore many methods have been researched so far to solve this problem.

[Thies et al. 2016] proposed a method which can reenact the video while maintaining source actor's visual-audio synchronization by using 3D facial statistical models. However their method cannot be applied to videos in which faces cannot be 3D-reconstructed, for example, vintage videos, 2D animations and more. On the other hand, image-based methods can be applied to a variety of videos. [Ezzat et al. 2002] proposed an image-based method to generate a speech animation. However, as phonemes correspond to mouth images in one-to-one in their model, they cannot consider coarticulation. [Bregler et al. 1997] proposed an alternative image-based approach by reusing frames in which mouth motion synchronizes with new audio. This approach can achieve coarticulation, but phoneme-matching tables are required when applying to dubbing videos.

In this paper, we propose an image-based method to automatically generate a variety of dubbing videos with visual-audio synchronization by frame-reusing without phoneme information. Contribution of our method is as follows. Our method 1) can automatically generate dubbing videos with visual-audio synchronization without any prior knowledges(e.g. phoneme information, 3D face models, generative models etc.) 2) can express coarticulation and 3) can be applied to a variety of videos such as mentioned above.

2 Our Method

We use two videos as input, one is the original video recording the actor's performance (the target video), and the other is that recording the voice-actor ' s. Our method mainly consists of " Frame-Reshuffling" part and "Mouth Region Composition" part. Fig.1 shows an outline of our method. In the "Frame-Reshuffling" part, first of all, the voice-actor ' s mouth shape is normalized to be close to the actor ' s. Then a noise-robust model to capture mouth shapes both in the target and in the source video is constructed by performing Principal Component Analysis(PCA). After that, we reshuffle frames in the target video for the target mouth motion to be the same with the source, then stitch and interpolate the selected frames to generate a new video. In the "Mouth Region Composition" part, we synthesize the mouth region in the new video to the original target video to maintain the original actor's motion.

2.1 Frame Reshuffling

(1)Matching Voice-Actor's Mouth Shape with Actor's

To calculate the similarity between the actor's mouth shape and the voice-actor's, normalizing their characteristics is required. Therefore we change the voice-actor's mouth shape to be closer to the actor's. We estimate N mouth feature points $\boldsymbol{S}_i^{tar}$ and $\boldsymbol{S}_i^{src}$, where i is the frame index of the videos and $tar(src)$ is the target(source) video. This time we use a method mentioned by [Irie et al. 2011]. Then we calculate the difference $\boldsymbol{\Delta}$ between $\boldsymbol{S}_0^{tar}$ and $\boldsymbol{S}_0^{src}$ and get normalized feature points $\boldsymbol{S}_i^{src'}$ by Eq.(1).

$$\boldsymbol{S}_i^{src'} = \boldsymbol{S}_i^{src} + \Delta = \boldsymbol{S}_i^{src} + \boldsymbol{S}_0^{tar} - \boldsymbol{S}_0^{src} \tag{1}$$

(2)Constructing a Noise-Robust Mouth Shape Model

We construct a noise-robust model by performing PCA to depict mouth shapes quantitatively on target mouth feature points $\boldsymbol{S}_i^{tar}$:

$$\boldsymbol{S}(\boldsymbol{\omega}) = \bar{\boldsymbol{S}}^{tar} + \boldsymbol{P}\boldsymbol{\omega} \tag{2}$$

where $\boldsymbol{S}$ is a mouth feature points vector, $\bar{\boldsymbol{S}}^{tar}$ is the average of $\boldsymbol{S}_i^{tar}$, $\boldsymbol{P}$ is the principal component matrix and $\boldsymbol{\omega}$ is a weight vector. Then we calculate each weight vector $\boldsymbol{\omega}_i^{tar}$ ($\boldsymbol{\omega}_i^{src'}$) from $\boldsymbol{S}_i^{tar}$ ($\boldsymbol{S}_i^{src'}$) using Eq.(2).

(3)Selecting Similar Mouth

In this process, we treat consecutive t frames as a unit. We select a target unit which has the most similar mouth shape to a source unit by minimizing Eq.(3) and continue this process, shifting the beginning frame index of a source unit, i, by $t-1$.

$$E_{i,j} = \begin{cases} \sum_{k=0}^{t} ||\boldsymbol{\omega}_k^{src'} - \boldsymbol{\omega}_{j+k}^{tar}||_2^2 \quad (i=0) \\ \alpha \sum_{k=0}^{t} ||\boldsymbol{\omega}_{i+k}^{src'} - \boldsymbol{\omega}_{j+k}^{tar}||_2^2 + (1-\alpha)||\boldsymbol{v}_j^{tar} - \boldsymbol{v}_l^{tar}||_2^2 \, (i>0) \end{cases} \tag{3}$$

SIGGRAPH '16, July 24-28, 2016, Anaheim, CA,
ISBN: 978-1-4503-4371-8/16/07
DOI: http://dx.doi.org/10.1145/2945078.2945097

Figure 2: *Ground Truth vs. Our Result*

In Eq.(3), j is the beginning frame index of a target frame unit. l is the end frame index of the target unit selected in the prior step. $\boldsymbol{v}_j^{tar}$ is a vector of estimated facial feature points in the jth target frame. α is a weight parameter to control the similarity of mouth shapes between the target and the source and the continuation of the target head motion. Note that when minimizing Eq.(3), each term is normalized and ranges from 0 to 1. Then we apply a frame-interpolation method for videos [Saito et al. 2014] to stitch together the selected target units seamlessly.

2.2 Mouth Region Composition

The video generated in Sec.2.1 does not maintain other motion (body motion, background motion etc.). Therefore we synthesize the mouth region to the original target video. First, we choose n undeformable facial feature points and track them through video frames and express them as $\boldsymbol{A}_i$ and $\boldsymbol{B}_i$, where A means the original target video, B means the video generated by "Frame-Reshuffling" part and i is the frame index.

$$\boldsymbol{A}_i = \begin{pmatrix} x_{i,1}^A & x_{i,2}^A & \dots & x_{i,n}^A \\ y_{i,1}^A & y_{i,2}^A & \dots & y_{i,n}^A \end{pmatrix} \tag{4}$$

$$\boldsymbol{B}_i = \begin{pmatrix} x_{i,1}^B & x_{i,2}^B & \dots & x_{i,n}^B \\ y_{i,1}^B & y_{i,2}^B & \dots & y_{i,n}^B \end{pmatrix} \tag{5}$$

We calculate a rotation matrix $\boldsymbol{R}_i$ and a translation vector $\boldsymbol{t}_i$ by minimizing Eq.(6) on each i based on Singular Value Decomposition method [Tamaki 2009]

$$\underset{\boldsymbol{R}_i, \boldsymbol{t}_i}{\arg\min} ||\boldsymbol{A}_i - (\boldsymbol{R}_i \boldsymbol{B}_i + \boldsymbol{t}_i)||_F^2 \tag{6}$$

Then we align the face in the generated video to that in the original target video and synthesize the mouth region by Poisson Image Editing[Pérez et al. 2003].

3 Result and Future Work

Fig.2 describes results of our method. This shows qualitatively that our method can generate dubbing videos with plausible visual-audio synchronization like the ground truth. In addition, we compared our results with traditional dubbing video by using RMSE as shown in Fig.3. Note that we scaled the RMSE values by regarding the length between inner corners of eyes as 30mm. It is clear that the RMSE values between our result and the ground truth are much smaller than that between a traditional dubbing video and the ground truth in almost all frames. From these results it can be said that our method can create mouth motion much more similar to

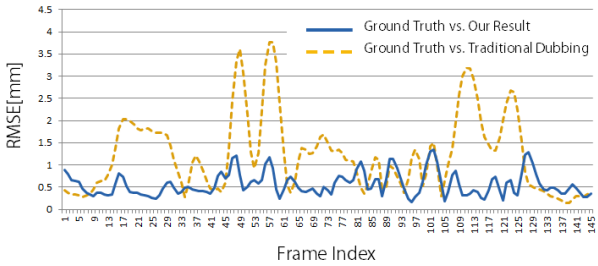

Figure 3: *Evaluation by RMSE*

ground truth. We also compared consecutive frames in a result and a ground-truth video and concluded that our method can consider coarticulation. This time, we applied our method to generate an English-dubbing video by a Japanese-spoken video and a Japanese-dubbing video from English-spoken. In both cases we confirmed our method worked well. This means our method can be applied independently of phonemes and suites for generating dubbing videos. As future work, we are focusing on improving our method to be robust to facial rotation by applying image-based rigid transformation and to illumination variation by illuminant estimation and image re-rendering.

References

BREGLER, C., COVELL, M., AND SLANEY, M. 1997. Video rewrite: Driving visual speech with audio. In *Proceedings of the 24th annual conference on Computer graphics and interactive techniques*, ACM Press/Addison-Wesley Publishing Co., 353–360.

EZZAT, T., GEIGER, G., AND POGGIO, T. 2002. *Trainable videorealistic speech animation*, vol. 21. ACM.

IRIE, A., TAKAGIWA, M., MORIYAMA, K., AND YAMASHITA, T. 2011. Improvements to facial contour detection by hierarchical fitting and regression. In *Pattern Recognition (ACPR), 2011 First Asian Conference on*, IEEE, 273–277.

PÉREZ, P., GANGNET, M., AND BLAKE, A. 2003. Poisson image editing. In *ACM Transactions on Graphics (TOG)*, vol. 22, ACM, 313–318.

SAITO, S., SAKAMOTO, R., AND MORISHIMA, S. 2014. Patchmove: Patch-based fast image interpolation with greedy bidirectional correspondence.

TAMAKI, T. 2009. Pose estimation and rotation matrices. *IEICE Technical Report. SIS 109*, 203, 59–64.

THIES, J., ZOLLHÖFER, M., STAMMINGER, M., THEOBALT, C., AND NIESSNER, M. 2016. Face2face: Real-time face capture and reenactment of rgb videos. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE, June 2016*.