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]

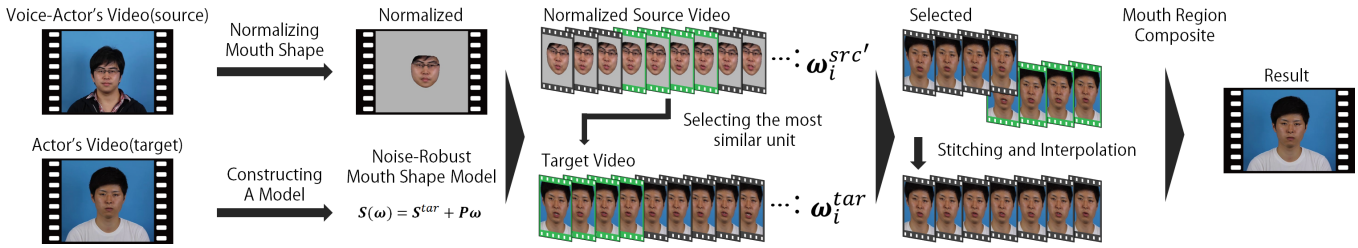

Figure 1: *Outline*

1 Introduction

Recently numerous videos have been translated into other languages using "dubbing" to make people enjoy the video contents without feeling language barriers. In general, dubbed videos have the advantages over subtitled ones that they do not draw the attention of the audience away from actors' performance. However, it is very difficult and complex to achieve the mouth-audio synchronization. That is to say a new dubbing audio does usually not synchronize with actor's mouth motion. This discrepancy can make the audience not only feel uncomfortable but also disturb comprehension of video contents. Therefore for high-quality dubbing, some methods to easily achieve mouth-audio synchronization are required.

Generally in dubbing process, translators translate the original sentences considering three time-consuming steps shown as follows.

1. The length of sentences: The duration of mouth motion must match with a new translated sentence.
2. The timing of breath: The timing of the actor's breath requires to synchronize with that followed by new sentences.
3. Mouth shapes: The actor's mouth shapes require to be similar to those pronouncing new sentences.

In spite of translators' great efforts, the new audio does usually not synchronize with actor's mouth motion. In addition, because of these restrictions, translators cannot feel free to translate into words they really want to use. Therefore to solve these problems, in this paper we focus on the method that directly modifies the actor's mouth motion to synchronize with the new dubbed audio.

2 Related Work

Many methods to generate realistic speech animation have been researched so far and they can be classified into two categories; one is based on 3D facial models and the other is based on image and audio processing.

2.1 Methods based on 3D facial models

[Weise et al. 2009] captured facial shapes by linear 3D blendshape models using RGB-D data. However, this method requires professional equipment to get RGB-D data. Then [Weise et al. 2011] presented a real-time facial expression capturing method using a commodity depth sensor. For improved fidelity, [Cao et al. 2014] combined depth input data with sparse facial features, and constructed personalized tracking model. These methods work well for capturing facial and transferring expressions, but they mainly focused on controlling 3D characters' facial expressions. On the other hands, [Thies et al. 2015] proposed a system to modify speaker's mouth motion in a video based on 3D facial models by commodity RGB-D sensor. Then, [Thies et al. 2016] achieved mouth motion transfer between two people using RGB data. These facial expression transfer methods are robust to 3D facial rotation, but cannot be applied to videos in which faces cannot be 3D-reconstructed, for example, vintage videos, 2D animations and more.

2.2 Methods based on image and audio processing

[Ezzat et al. 2002] demonstrated an image-based method to generate a realistic speech animation. However they require to construct a different generative model per person. [Chang and Ezzat 2005] refine the generative model constructed from one person in order to be applied to others by taking correspondence between people. However, in these models, as phonemes correspond to mouth images in one-to-one, they cannot consider coarticulation. On the other hand, [Bregler et al. 1997] modified mouth motion in videos based on 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 multi-languages like dubbing.

2.3 Our Contribution

In this paper, we propose an image-based method to automatically generate dubbed videos with mouth-audio synchronization by frame-reshuffling without phoneme information, 3D facial models and more. Contribution of our method is as follows. Our method 1) can automatically generate dubbed videos with mouth-audio synchronization without any prior knowledges (e.g. phoneme information, 3D facial statistical models, image-generative models etc.) 2) can express coarticulation and 3) can be applied to a variety of videos including un-3Dreconstructable videos and 2D animations.

3 Our Method

We use two videos as input, one is the original video capturing actor's performance (target video), and the other is that capturing voice-actor's performance (source video). Our key idea is to achieve mouth-audio synchronization by shuffling target video frames for the actor's mouth motion to appear at the same timing with the voice-actor's one. Our method mainly consists of three steps; 1)"Pre-Processing" part, 2)"Frame-Reshuffling" part and 3)"Mouth Region Composition" part. Fig.1 shows the overview of our method. In the "Pre-Processing" part, at first we normalize the voice-actor's mouth shape to be close to the actor's. Then

a noise-robust model to capture mouth shapes quantitatively is constructed by performing principal component analysis (PCA) to mouth shape landmarks of the actor. After that, in the "Frame-Reshuffling" part, we reshuffle the target video frames as described above, and then stitch and interpolate the reshuffled frames to generate a more seamless video (reshuffled video). Then in the "Mouth Region Composition" part, we synthesize the mouth region in the reshuffled video to the original target video to maintain other motion (e.g. the actor's body motion, background motion etc.).

3.1 Pre-Processing

(1)Making voice-actor's mouth shape close to actor's one
In general, mouth shape is different from each other of people. Therefore to calculate the similarity between the actor's mouth motion and the voice-actor's one, we require to normalize their characteristics. Specifically, we change the voice-actor's mouth shape to be closer to the actor's one. First of all, we detect $N = 22$ mouth feature landmarks $\boldsymbol{S}_i^{tar}$ and $\boldsymbol{S}_i^{src}$ using the method proposed by [Irie et al. 2011] (Fig.2), where i is the frame index of the videos and $tar(src)$ means the target (source) video. Note that after detecting landmarks, we apply Gaussian filter, whose window-size is 5 frames, to the landmark coordinates to smooth them. Then we calculate the difference $\boldsymbol{\Delta}$ between $\boldsymbol{S}_0^{tar}$ and $\boldsymbol{S}_0^{src}$, and get normalized feature landmarks $\boldsymbol{S}_i^{src'}$ by Eq.(1).

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

(2)Constructing a noise-robust mouth shape model
To depict mouth shapes quantitatively, we construct a noise-robust model by performing PCA to the target mouth feature landmarks $\boldsymbol{S}_i^{tar}$ and get Eq.(2).

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

where each character is as follows

$\boldsymbol{S} = (x_1, y_1, \cdots, x_N, y_N)^T$:a mouth feature landmarks vector
$\bar{\boldsymbol{S}}^{tar} = \sum_{i=1}^{L_{tar}} \boldsymbol{S}_i^{tar}$:the average of $\boldsymbol{S}_i^{tar}$ (L_{tar} is the length of target video)
$\boldsymbol{P} = (\boldsymbol{p}_1, \boldsymbol{p}_2, \cdots, \boldsymbol{p}_M)$:the principal component matrix
$\boldsymbol{\omega} = (\omega_1, \omega_2, \cdots, \omega_M)^T$:a weight vector

M is the number of the principal component vectors and in our experiment, we use $M = 16$. Then in the next step we calculate the mouth shape similarity between the actor and the voice-actor based on the weight vectors $\boldsymbol{\omega}_i^{tar}$ and $\boldsymbol{\omega}_i^{src'}$ calculated from $\boldsymbol{S}_i^{tar}$ and $\boldsymbol{S}_i^{src'}$ using Eq.(2).

3.2 Frame Reshuffling

Fig.3 shows the outline of "Frame Reshuffling" process.
(1)Selecting Similar Mouth Shapes
In this process, we regard consecutive t frames as a unit. We define the similarity of mouth shapes between a target unit and a source unit as Eq.(3).

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

where $||\cdot||_2$ is the L_2-norm, i is the beginning frame index of a source unit and j is the beginning frame index of a target unit.

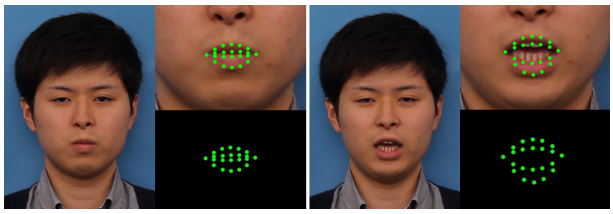

Figure 2: *Mouth Feature Landmarks.*

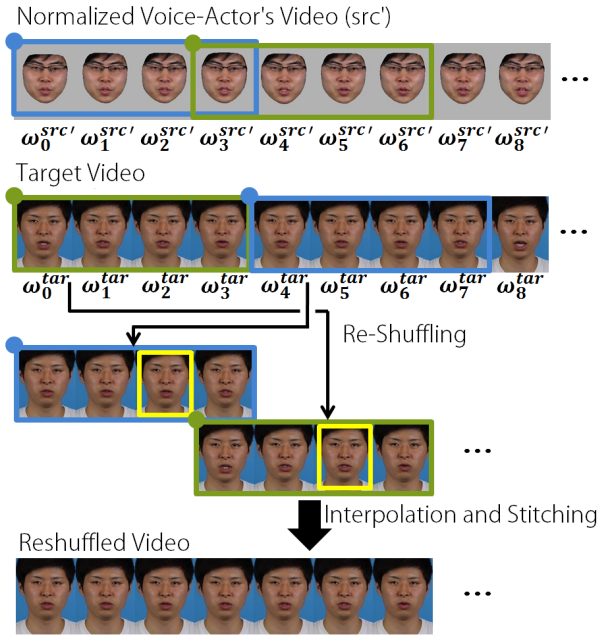

Figure 3: *Outline of shuffling and stitching video frames.*

When $i = 0$, we select a target unit with the most similar mouth shapes to the source unit, which consists of the $0 \leq k \leq t-1$th source frame, by minimizing Eq.(3) ($i = 0$). Then we update i by shifting it by $t-1$, where $t-1$ means allowing overlap of the units by one frame, and select a similar target unit by minimizing Eq.(3) ($i > 0$). Note that the second term on the right hand of Eq.(3) ($i > 0$) shows the similarity of facial landmark coordinates detected by using [Irie et al. 2011] ($\boldsymbol{v}_j^{tar}$ is the facial landmark vector of the jth target frame and l is the end frame index of the similar target unit selected in the prior step $i = 0$). To add this term, we achieve selecting a more seamless unit to that selected in the prior step. α is a weight parameter to control the first term and the second one. Note that when minimizing Eq.(3), each term is normalized and ranges from 0 to 1. Then we continue to update i and generate a reshuffled target video.

(2)Interpolating and stitching frame units
To stitch together the selected target units seamlessly, we require to interpolate the overlapping frames. This time, we apply a frame interpolation method for videos [Saito et al. 2014]. In detail, we generate two frames located at the seam where two units are stitched (the two frames between the yellow-marked frames in Fig.3). After performing this process at all seams, the final reshuffled video is generated.

3.3 Mouth Region Composition

The actor's mouth motion in the video generated in Sec.3.2 synchronizes with the voice-actor's one. However, it does not maintain

other motion (e.g. body motion, background motion etc.). Therefore, to solve this problem, we synthesize the mouth region of the reshuffled video to the original target one. First, we select $n = 6$ undeformable facial feature landmarks (inner corners of eyes, three contour points of nose and tip of nose) and track them through video frames, then expressing them as $\boldsymbol{A}_i$ and $\boldsymbol{B}_i$ like Eq.(4)(5), where $\boldsymbol{A}$ means the original target video, $\boldsymbol{B}$ means the reshuffled video and i is the frame index.

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

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

Then using $\boldsymbol{A}_i$ and $\boldsymbol{B}_i$, we calculate a rotation matrix $\boldsymbol{R}_i$ and a translation vector $\boldsymbol{t}_i$ by minimizing Eq.(6) based on Singular Value Decomposition method[Tamaki 2009].

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

where $||\cdot||_F$ is the Frobenius-norm and we assume that $\boldsymbol{A}_i$ and $\boldsymbol{B}_i$ have the same scale because the reshuffled video is generated by shuffling the original target video. Then we align the actor's face in the reshuffled video to the original target video and synthesize the mouth region by Poisson Image Editing[Pérez et al. 2003].

4 Result

We applied our method to generate a Japanese-dubbed video from an English one. Fig.7 compares our result with a ground truth video and a traditional dubbing video. Here, in the ground truth, the actor is speaking a Japanese sentence and traditional dubbing is one of the methods usually performed by translators, in which translators simply overlap voice-actor's voice over actor's videos. This time, the actor is speaking about 5 sets of "Harvard Sentences" (about 5 minutes, 29.97fps) and the voice-actor is speaking a sentence translated from arbitrarily-selected one of the sentences spoken by the actor (about 4 seconds, 29.97fps, speech rate:8.37mora/sec). Here, mora is a sound unit of speech, and for example, "ko-n-ni-chi-wa" consists of 5 mora. Then we used $\alpha = 0.75$ and $t = 6$ in Eq.(3). Fig.7 shows that our method can generate a dubbed video with plausible mouth-audio synchronization; in fact, our result is much more similar to the ground truth compared with the traditional dubbing. Furthermore, in order to evaluate coarticulation, we also compare the consecutive frames in our result and the ground truth as shown in Fig.6. From Fig.6, it is confirmed that the mouth motion in our result is similar to the ground truth even in a short period, which concludes that our method can consider the coarticulation. We also applied our method to generate an English-dubbed video from a Japanese-spoken video. This time, the Japanese actor is speaking 1 set of "ATR 503 sentences"(about 5 minutes, 29.97fps) and the voice-actor is speaking an arbitrary English sentence (about 3 seconds, 29.97fps, speech rate:3.13 words/sec). Used parameters in Eq.(3) are as follows; $\alpha = 0.75$ and $t = 8$. Fig.8 shows the comparison with ground truth and traditional dubbing. From Fig:8, we can also confirm our result is much similar to the ground truth. Therefore, it is clear that our method worked well for generating dubbed videos independently of languages.

In addition, we also qualitatively compare our result with the ground truth and the traditional dubbing by RMSE(Fig.4,5). Note that we scaled the distance between inner corners of eyes as 30mm when calculating RMSE values. It is clear that in almost all frames, the RMSE values calculated from our result and the ground truth are much smaller than those from the ground truth and the traditional

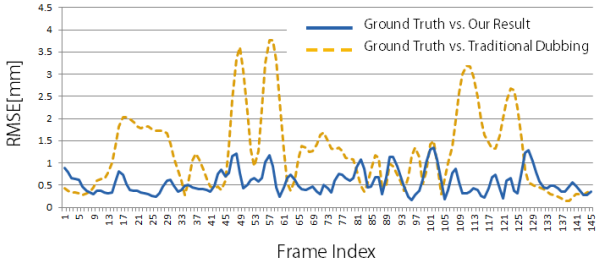

Figure 4: *Evaluation by RMSE(English to Japanese)*

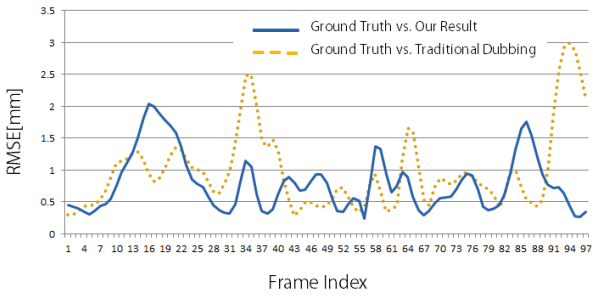

Figure 5: *Evaluation by RMSE(Japanese to English)*

dubbing. In the beginning of Fig.5, the rmse gets high values and this is because here is breathing period and while the actor breathed shortly with the mouth narrowly open, the voice-actor breathed with his mouth widely open. In conclusion, we confirm our result works well quantitatively.

5 Conclusion and Future Work

We propose an approach to solve the mouth-audio discrepancy in dubbed videos based on "Frame Reshuffling". Our key idea is to shuffle an actor's video frames for actor's mouth motion to appear at the same timing with voice-actor's and this approach achieves to generate dubbed videos with mouth-audio synchronization without using any prior knowledge (e.g. phoneme information, 3D facial models, generative models etc.). In addition, as we treat some consecutive video frames as a unit, which maintain natural speaking mouth motion, our method can consider coarticulation. Moreover, as our method is image-based, it can be applied to a variety of videos (e.g. un-3Dreconstructable videos, 2D cartoon animation etc.). Then this time, we evaluate our result both qualitatively and quantitatively and conclude that our method can generate plausible mouth motion similar to ground truth independently of languages.

In our method, we use sparse facial landmarks in order to align mouth region, so when applying our method to videos including large actor's head rotation, some artifacts of image synthesis will occur. Therefore as future work, we focus on improving robustness to head rotation by pixel-based alignment methods. This time, we applied our method to videos captured under constant illumination, so when focusing on videos with scene changes, we require to apply our method scene by scene. In order to solve this problem, we plan to automatically segment scenes and improve the applicability of our method.

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*

Figure 7: *Comparison between a Japanese result generated from an English video and ground-truth/a traditional dubbing video.*

Figure 6: *Coarticulation.*

techniques, ACM Press/Addison-Wesley Publishing Co., 353–360.

CAO, C., HOU, Q., AND ZHOU, K. 2014. Displaced dynamic expression regression for real-time facial tracking and animation. *ACM Transactions on Graphics (TOG) 33*, 4, 43.

CHANG, Y.-J., AND EZZAT, T. 2005. Transferable videorealistic speech animation. In *Proceedings of the 2005 ACM SIGGRAPH/Eurographics symposium on Computer animation*, ACM, 143–151.

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 bidirec-

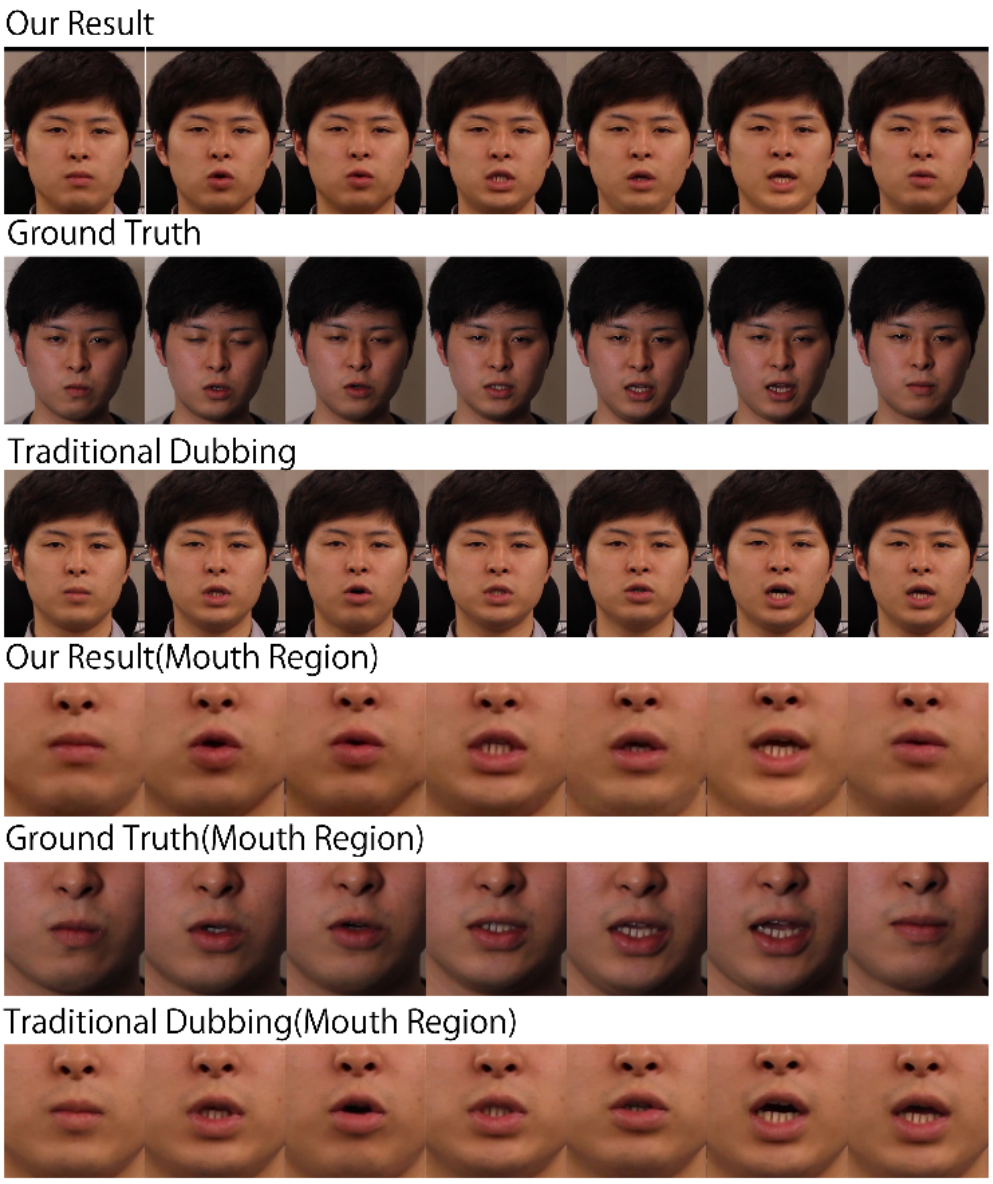

Figure 8: *Comparison between an English result generated from a Japanese video and ground-truth/a traditional dubbing video.*

tional correspondence.

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

THIES, J., ZOLLHÖFER, M., NIESSNER, M., VALGAERTS, L., STAMMINGER, M., AND THEOBALT, C. 2015. Real-time expression transfer for facial reenactment. *ACM Transactions on Graphics (TOG) 34*, 6.

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*.

WEISE, T., LI, H., VAN GOOL, L., AND PAULY, M. 2009. Face/off: Live facial puppetry. In *Proceedings of the 2009 ACM SIGGRAPH/eurographics symposium on computer animation*, ACM, 7–16.

WEISE, T., BOUAZIZ, S., LI, H., AND PAULY, M. 2011. Realtime performance-based facial animation. In *ACM Transactions on Graphics (TOG)*, vol. 30, ACM, 77.