EUROGRAPHICS 2015 / O. Sorkine-Hornung and M. Wimmer
(Guest Editors)

Volume 34 (2015), Number 2

VDub: Modifying Face Video of Actors for Plausible Visual Alignment to a Dubbed Audio Track (Supplementary Material)

P. Garrido [†1] and L. Valgaerts [‡1] and H. Sarmadi[1] and I. Steiner[2] and K. Varanasi[3] and P. Pérez[3] and C. Theobalt[1]

[1]MPI for Informatics [2]Saarland University and DFKI GmbH [3]Technicolor

Abstract

This document provides a more in-depth explanation of our experimental validations, described in Sec. 7.2 in the main paper. They include more details concerning the user study and the statistics obtained from it, as well as an extensive description of how we performed the comparison to the image-based approach presented in the paper. We also provide screenshots of the comparison shown in the supplementary video.

1. User Study

To perform our user study, we created a webpage showing the results for the three sequences mentioned in the paper, namely the recited passage of a movie (Fig. 7), the scene of a passion play (Fig. 8), and the interview sequence (Fig. 9). Each result was compared to that obtained by traditional dubbing, side by side in a random order. The traditionally dubbed results were provided by the same dubbing studio that recorded the dubbing actor and dubbed the German language track into English. These videos were generated by taking the original German target videos, removing the original audio, and adding the dubbed language track in English. Note that the dubbed language track was further altered (i.e., manually time-shifted and skewed) by one of the experts in the dubbing studio to improve the overall audio-visual alignment, thus creating high-quality professional videos (please refer to the additional supplementary video for these results). The results corresponding to the same sequence were equally long and their lengths, as well as other features, such as the amount of head motion, can be seen in Table 1.

To quantify the quality of the dubbing, we attached to the results a questionnaire that evaluated the overall audio-visual experience, including viewing discomfort and how natural the video-audio combination was perceived by the user overall. To be precise, we included a Likert scale that ranged from 0 to 5, where 0 means a really bad overall visual-audio experience, while 5 means a very good experience. To collect some statistics about the user preference, we also asked the users to give their preference for one of the two approaches. An optional comment box was also included.

The web-based user study was sent to 45 participants from different places around the world, including countries were dubbing is a common practice (Germany and France) and also countries were it is not (UK, USA and Chile). All the participants were required to understand and speak English well; German was not required at all, since none of the results were displayed with the original German audio track.

1.1. Results and Statistics

Table 2 summarizes the overall scores assigned to each sequence, as well as user preferences.

We additionally performed the ANOVA F-test to find the statistical significance of the scores obtained in our user study. The p-values were $\sim$0.4, 0.001, and 0.006 for the results of Fig. 7, Fig. 8, and Fig. 9, respectively. This means that two out of three experiments were statistically significant, as their p-value falls below 0.01, i.e., the random sampling error in the user study is less than 1%. The high p-value of the experiment related to the result of Fig. 7 can be ascribed to the high standard deviations and the tied scores compared to the others, meaning that more samples would be needed to have conclusive statistics. However, we believe that the scores for this sequence illustrate a trend towards equal appreciation of our results and those of the studio.

† e-mail: pgarrido@mpi-inf.mpg.de
‡ e-mail: valgaerts@mpi-inf.mpg.de

Table 1: *Length and main features of the sequences used in the user study.*

Sequence	length (sec)	head motion	head orientation
Movie dialog (Fig. 7)	30	negligible	frontal
Passion play (Fig. 8)	20	strong/fast	frontal/non-frontal
Interview (Fig. 9)	30	mild	mostly non-frontal

Table 2: *Scores given by the survey respondents to the results attained by traditional dubbing and our approach, as well as their overall preference.*

Sequence	Traditional dubbing		Our approach	
	Score	Preference	Score	Preference
Movie dialog (Fig 7)	2.9 ± 1.2	53%	2.7 ± 1.1	47%
Passion play (Fig 8)	3.0 ± 1.2	73%	2.3 ± 0.9	27%
Interview (Fig. 9)	3.6 ± 1.1	70%	3.0 ± 0.9	30%
Overall	3.2 ± 1.2	65%	2.7 ± 1.0	35%

In general, the variability of the scores can be explained by two main criteria that the users found very relevant: lip-sync and expressiveness / realism. Some survey respondents preferred good lip-sync to out-of-sync but expressive faces, but also the other way round. Some of the comments left by the participants include: "Sometimes exaggerated expression is better", "I voted for the videos where the sound-image synchronization was better", "In my opinion, it is not just about making the mouth move in line with the audio", "I feel that the synchronization by itself always looked very good".

2. Comparison to Image-Based Approaches

To demonstrate that our 3D model-based approach outperforms 2D image-based approaches, we compare VDub to a modified version which does not produce the final composites by rendering a synthesized 3D geometry, but by reordering the target frames and applying non-rigid 2D warping as in the *face reenactment* approach of [GVR*14]. Such a method is similar to a purely image-based technique, like Video Rewrite [BCS97], but with better image warping.

In our dubbing scenario, the *face reenactment* approach of [GVR*14] would replace the dubber's face in the dubbing sequence with the face of the actor, while preserving the original dubber's performance as much as possible. To do so, the method consists of three main steps: face tracking, face matching, and face transfer. Face tracking accurately tracks sparse facial landmarks. Face matching encodes the facial landmarks as histograms of local binary patterns and uses them to match target and dubbing frames to produce a reordering of the original target sequence. Face transfer then warps and blends the face region of the reordered target frames back into the dubbing sequence using the tracked landmarks, and a combined global and local non-rigid 2D mapping. The original method of [GVR*14] therefore creates a new synthesized sequence with a facial performance that is close to that of the dubber, but it warps the actor into the dubbing sequence (instead of the original target sequence) and mixes the identities of the dubber and the actor, which is not suitable for our dubbing scenario.

We can design an image-based approach that is suitable for our dubbing scenario by reordering the target frames as in [GVR*14], but warping the face region of the retrieved target frames back into the original target sequence. To assist the warping, we use the synthesized facial landmarks that are provided by the motion transfer step in VDub. These landmarks correspond to the actor's face in the target sequence, but move in accordance to the dubber's speech. For the warping, we can use the same non-rigid 2D mapping defined in [GVR*14]. The resulting strategy is image-based and ensures that the mouth motion of the warped frames moves in pace with the dubber's mouth motion while being correctly aligned to the actor's face in the original target sequence.

2.1. Results

Figure 1 shows some of the results obtained by VDub and by the image-based approach on the sequence of Fig. 8 of the paper. Note that the image-based approach replaces the complete inner face, as in the method of [GVR*14], while our method only replaces the lower part of the face. The image-based results can suffer from ghosting artifacts (third column), may not always be in pace with the dubber's performance (second and fifth column) and may struggle with strong head motion and unrealistic face warping (third column). These and other issues, such as the temporal alignment of the mouth region and the temporal resolution, can be more clearly seen in the supplementary video. This demonstrates that our model-based approach produces synthesized sequences of overall higher quality in terms of the spatio-temporal resolution and can deal well with challenging sequences that exhibit fast and strong head motion, where image-based approaches normally have trouble.

Figure 1: *Final composite using our model-based approach (top), final composite obtained by the image-based approach (middle), and the corresponding frames from the dubbing sequence (bottom).*

References

[BCS97] BREGLER C., COVELL M., SLANEY M.: Video Rewrite: Driving visual speech with audio. In *ACM TOG (Proc. SIGGRAPH)* (1997), pp. 353–360. 2

[GVR*14] GARRIDO P., VALGAERTS L., REHMSEN O., THORMAEHLEN T., PEREZ P., THEOBALT C.: Automatic face reenactment. In *Proc. CVPR* (2014). 2