Automatic Face Reenactment
Supplementary Material

Pablo Garrido[1] Levi Valgaerts[1] Ole Rehmsen[1] Thorsten Thormählen[2]
Patrick Pérez[3] Christian Theobalt[1]
[1]MPI for Informatics [2]Philipps-Universität Marburg [3]Technicolor

Abstract

This document extends the main paper by presenting a more in-depth analysis of the temporal clustering approach proposed in Sec. 5.2 in the paper, expanding the validation paragraph of Sec. 7 in the paper, and presenting screenshots for the results shown in the supplementary video.

1. Temporal Clustering

Algorithm. Sec. 5.2 in the main paper introduces a temporal clustering approach that divides the target sequence into consecutive clusters of variable length, based on the appearance similarity of facial features between subsequent frames. Our clustering idea is related to hierarchical agglomerative clustering approaches, but it is explicitly designed to preserve temporal continuity, i.e., it only merges clusters that are consecutive in time, thereby preserving the order of the target frames. The approach is based on a distance metric for facial appearance between consecutive frames, defined as d_{app} in Eq. (3) in the main paper. As a linkage criterion, this distance metric is extended to two consecutive clusters $\mathcal{C}^{k-1}$ and $\mathcal{C}^k$ and it is computed as the average of pairwise distances between all frames in $\mathcal{C}^{k-1}$ and $\mathcal{C}^k$.

Our temporal clustering approach assumes that all frames represent single clusters to begin with. The approach then proceeds iteratively by merging only two clusters in each step, namely those that are currently the closest according to the distance metric d_{app}. The criterion for merging two clusters is that the variance of d_{app} between consecutive frames within the newly generated cluster can not become larger than the maximum variance of d_{app} between consecutive frames in the original clusters. This merging criterion ensures that frames within a cluster are as similar as possible. The algorithm terminates once this criterion is not met. We impose that merging always takes place if one of the two clusters that will be merged next contains a single frame. An advantage of our clustering approach is that it is parameter-free, so no tuning is required. The result is a sequence of target sections $\mathcal{C}^k$, with index k running in temporal direction over the total number of computed clusters.

Analysis. Fig. 1 shows a plot of the distance metric d_{app} between two consecutive frames for 32 frames of the target sequence depicted in Fig. 5. The target clusters that are computed by our temporal clustering approach are drawn as red lines below the graph, while isolated frames and boundary frames are indicated by green squares. The values of the distance metric d_{app} are drawn as red circles enclosed by the frames between which it measures the similarity.

As one would expect, consecutive frames are merged into a cluster if the value of d_{app} is low. If d_{app} remains low for an extended number of consecutive frames, a large cluster is formed, such as the one spanning frames 48 to 52. Peaks in the graph indicate dissimilar frames and these typically form cluster boundaries or isolated frames. Note that the graph is dynamic and changes as the algorithm proceeds since the value of d_{app} between consecutive clusters changes as more clusters are formed (difficult to visualize).

To illustrate the similarity in appearance of frames within the same cluster, we display the boundary frames of the cluster spanning frames 38 to 41 at the bottom left, the cluster spanning frames 48 to 52 in the top middle, and the cluster spanning frames 53 to 55 at the bottom right of the figure. The two examples of isolated frames shown at the top left and right side lie outside of a cluster and differ in appearance from those within the neighboring clusters. It can be concluded that the length of a cluster roughly varies inversely proportionally to the change in expression and the timing of speech within the cluster. The maximum and average cluster length and the total number of clusters computed for the target sequences of the figures below are given in Tab. 1. For the results presented here, we enforced the minimum cluster size to be 2, which generally leads to smoother animations for sequences with many isolated frames. Enforcing this is easily done by adding isolated frames to the

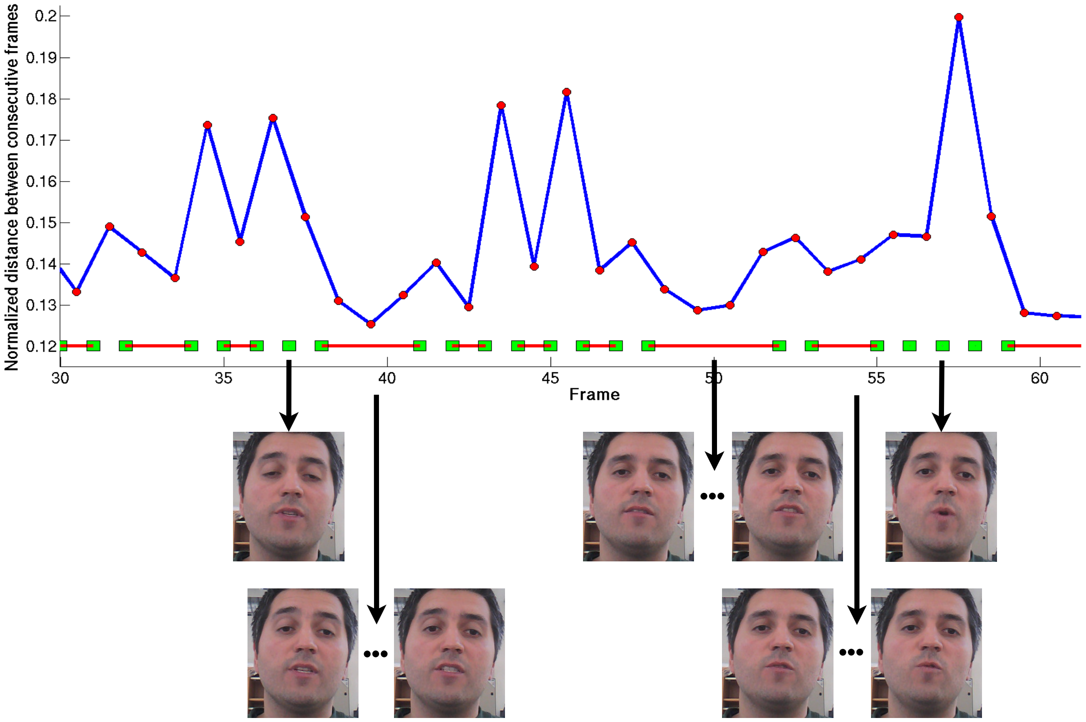

Figure 1. Plot of the distance metric d_{app} and the resulting clusters obtained by our temporal clustering approach.

Table 1. The mean (mean size) and maximum (max. size) cluster size, and the total number of clusters (num. clusters) computed for the sequences shown in this document.

Sequence	num. frames	num. clusters	mean size	max. size
Fig. 2	231	86	2.7	5
Fig. 3	524	196	2.7	6
Fig. 4	374	136	2.8	6
Fig. 5	200	59	3.5	15
Fig. 6	446	155	2.9	8
Fig. 7	566	215	2.6	6
Fig. 8	352	136	2.6	4
Fig. 9	319	119	2.7	6
Fig. 10	319	128	2.5	5
Fig. 12	533	191	2.8	9

left or right cluster, depending on which one is closest in d_{app}.

2. Experimental Validation

User study. We evaluated the different contributions of our system by comparing our full reenactment system with (1) a simplified system that does not include the temporal clustering approach proposed in Sec. 5.2 in the main paper (i.e., a straightforward frame-by-frame matching) and (2) a basic system that does not include temporal clustering, nor the motion distance defined in Eq. (4) in the main paper (i.e., a frame-by-frame matching which does not enforce temporally-coherent motion of landmarks). To this end, we performed a user study with 32 participants. The participants were asked to rate reenactment results for two low-quality (LQ) web videos and five existing high-quality (HQ) videos with respect to the original target performance in terms of mimicking fidelity, temporal consistency and visual artifacts on a scale from 1 (not good) to 5 (good). The study was conducted as a web page with the resulting videos that was presented to a general audience of non-experts that were not aware of the techniques employed to generate the reenactments. Tab. 2 shows the average rating for the seven sequences shown below, which also appear in the supplementary video. From these results, we conclude that our full system (3.25 average over all sequences) outperforms systems without temporal clustering (2.92), and additionally without combined appearance and motion distance (1.48). These results are statistically significant as the ANOVA p-value for each sequence was on average below 10^{-5}. Overall, the scores for the HQ sequences were higher than for the LQ web videos. These scores should not be directly compared to those reported by Li *et al*. [2], as we evaluated different methods and asked different questions.

Table 2. Results of a user study with 32 participants and seven of our reenactment results. The scores listed below denote the average of a rating between 1 (not good) and 5 (good) with respect to the original target performance in terms of mimicking fidelity, temporal consistency and visual artifacts. The results used in the study are the ones referred to by the figure number and are shown later in this document.

Sequence	LQ video		HQ video				
	Fig. 2	Fig. 3	Fig. 6	Fig. 7	Fig. 8	Fig. 9	Fig. 10
Our full system	**2.5**	**3.56**	**3.19**	**3.00**	**3.38**	**3.29**	**3.81**
without temporal clustering	2.09	2.84	3.16	2.72	3.06	3.13	3.47
without temporal clustering and motion distance	1.34	1.34	1.41	1.16	1.50	1.45	2.16

Self-reenactment. The supplementary video and Fig. 11 show an example of a self-reenactment, i.e. a reenactment result obtained by taking the same video sequence, both as source and target. Ideally, such a result should be identical to the input videos, and it can be used to test the performance of a reenactment system, for instance, by examining visual artifacts that are introduced in the original sequence.

The self-reenactment shown in Fig. 11 is almost indistinguishable in appearance and expression from the source and target video. If we define a mismatch as a source frame that is assigned to a target cluster in which it is not contained (source and target are the same video), our system produced 36 mismatches on a total of 214 clusters (22 s of video). The first two columns in Fig. 11 show two of such mismatches, where a cluster that appears earlier in the sequence was matched to a later frame. However, as it can be observed, these mismatches are very similar in appearance to the frames in the target clusters and the final reenactment is visually close to a perfect frame-by-frame synthesis of the true target sequence. This similarity is confirmed by an average PSNR of 41 dB over 566 frames, with a minimum of 33 dB. Fig. 5 shows a self-reenactment of a low-quality 10 s webcam sequence. We obtained 1 mismatch on 59 computed clusters.

Also for the case where source and target depict the same person under similar conditions, the reenactment resembles the target sequence closely. An example is shown in Fig. 10, where the source and target sequence are different excerpts taken from a 100 s recording of the same person. Both excerpts were selected arbitrarily without considering possible similarities in the actor's performance. The figure and the supplementary video show that the final reenactment is very convincing and realistic, a result that was also highly appreciated in the user study of Tab. 2 (last column).

Length of the source video and reenactment quality. To demonstrate the influence of the source data size on the reenactment quality, we repeated our experiments for successively shorter source sequences, i.e. by taking the first 50%, 25%, and 12.5% of the source material. The supplementary video shows such a test for the self-reenactment of Fig. 11. We conclude that a small amount of source frames may lead to unnatural results, with static expressions that appear to be stuck on a moving face (due to certain frames being selected repeatedly). Longer source sequences clearly result in more realistically reenacted expressions and fewer abrupt transitions, since the newly included source frames cover more of the expressions in the target sequence. However, for many of our other examples, the deterioration in reenactment quality with increasingly shorter source sequences was not as pronounced. This shows that we can even produce plausible results for a small set of source frames.

A near-perfect reenactment could be achieved for any target sequence by using a huge amount of meticulously preselected source frames that span a large dictionary of possible expressions. However, such results would strongly depend on the choice of database, and a main contribution of our paper is to demonstrate that our method works for videos containing arbitrary facial expressions.

Comparison with Dale *et al.* Finally, we compared our fully automatic reenactment system with the semi-automatic face replacement system of Dale *et al.* [1] on data provided by the authors. The source and target sequences depict two different subjects reciting the same poem. Our reenactment result is shown in Fig. 12 and in a side-by-side comparison with the result of Dale *et al.* in the supplementary video, demonstrating that they are visually very close in quality. A direct frame-by-frame comparison of both results is not meaningful since the method of Dale *et al.* transfers the source face, including the complete source performance, while our method only transfers the source face, but preserves the target performance. Because source and target performance for this example are slightly different (due to the poem being recited by two different actors), both results differ visually as well. Strictly speaking, the result of Dale *et al.* is not a "reenactment" as defined in our main paper: Their method warps the target timeline to match that of the source performance and transfers the source face, including its complete performance, which may be considered an easier task since it inherently ensures temporal continuity in the final composite.

3. Results

In the remainder of this document, we show some screenshots of the results attained by our system on low-quality Internet videos and high-quality existing videos which were utilized in the user study and presented in the main paper, as well as in the supplementary video.

References

[1] K. Dale, K. Sunkavalli, M. K. Johnson, D. Vlasic, W. Matusik, and H. Pfister. Video face replacement. *ToG (Proc. SIGGRAPH Asia)*, 30(6):130:1–130:10, 2011.

[2] K. Li, F. Xu, J. Wang, Q. Dai, and Y. Liu. A data-driven approach for facial expression synthesis in video. In *Proc. CVPR*, pages 57–64, 2012.

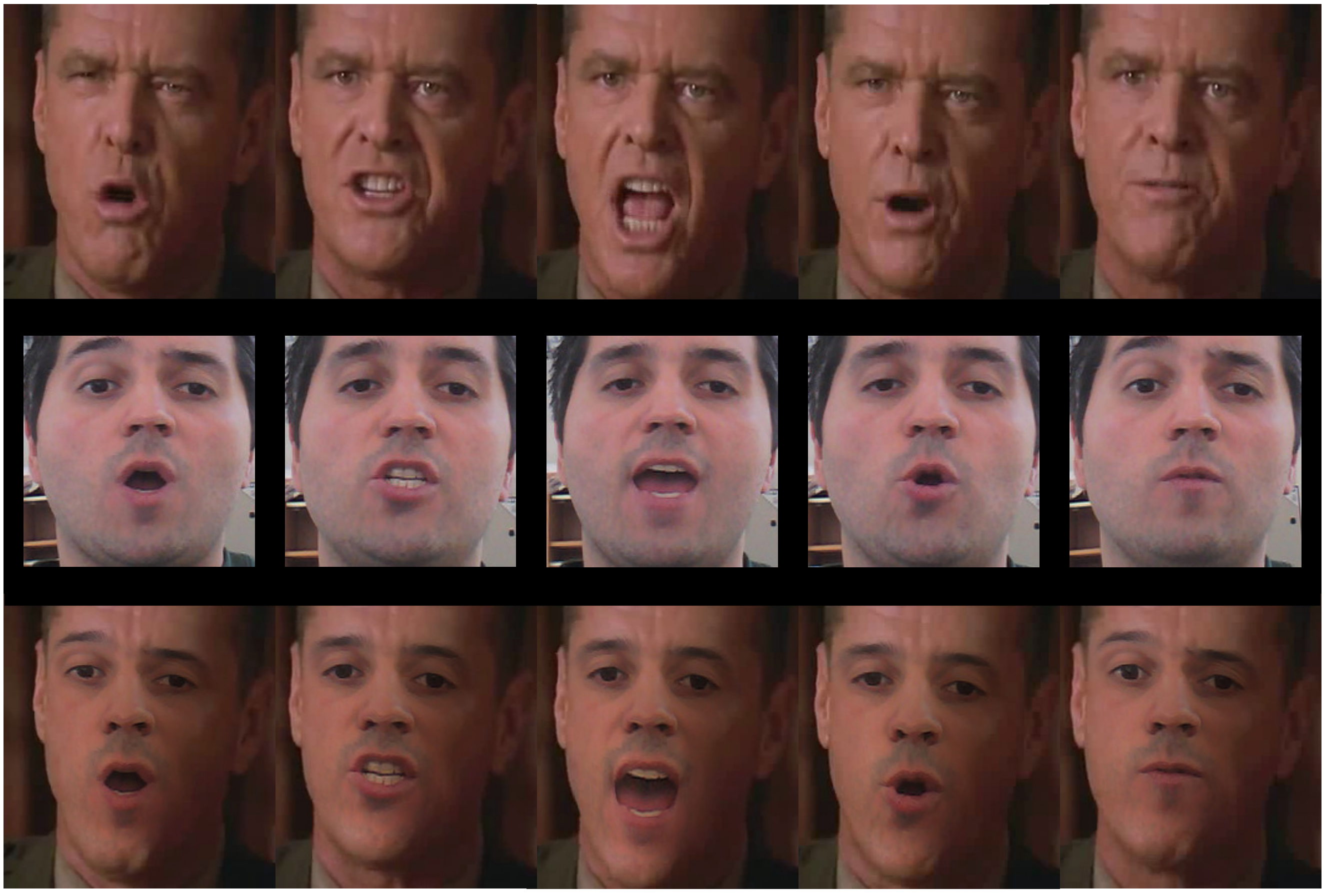

Figure 2. Low-quality video from the Internet (8 s of target footage, 10 s of source footage). Excerpt from "A Few Good Men" (http://youtu.be/5j2F4VcBmeo). Top: Frames from the target sequence. Middle: Corresponding selected source frames. Bottom: Final composites. Chosen weights in Eq. (9) in the main paper: $w_{\text{nr}} = 0.45$, $w_{\text{r}} = 0.55$.

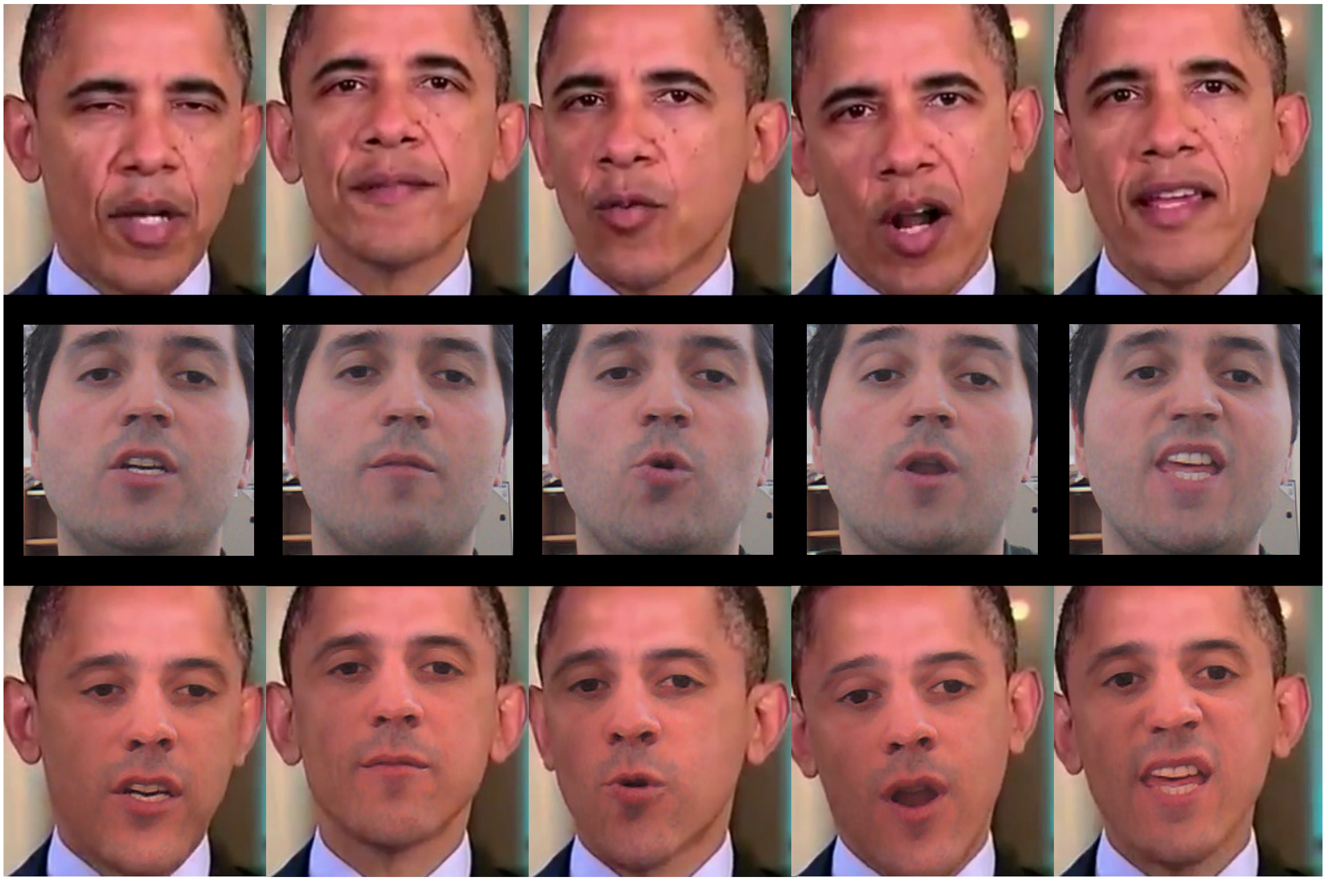

Figure 3. Low-quality video from the Internet (8 s of target footage, 10 s of source footage). President Obama's speech (http://youtu.be/qxtydXN3f1U). Top: Frames from the target sequence. Middle: Corresponding selected source frames. Bottom: Final composites. Chosen weights in Eq. (9) in the main paper: $w_{\text{nr}} = 0.65$, $w_{\text{r}} = 0.35$.

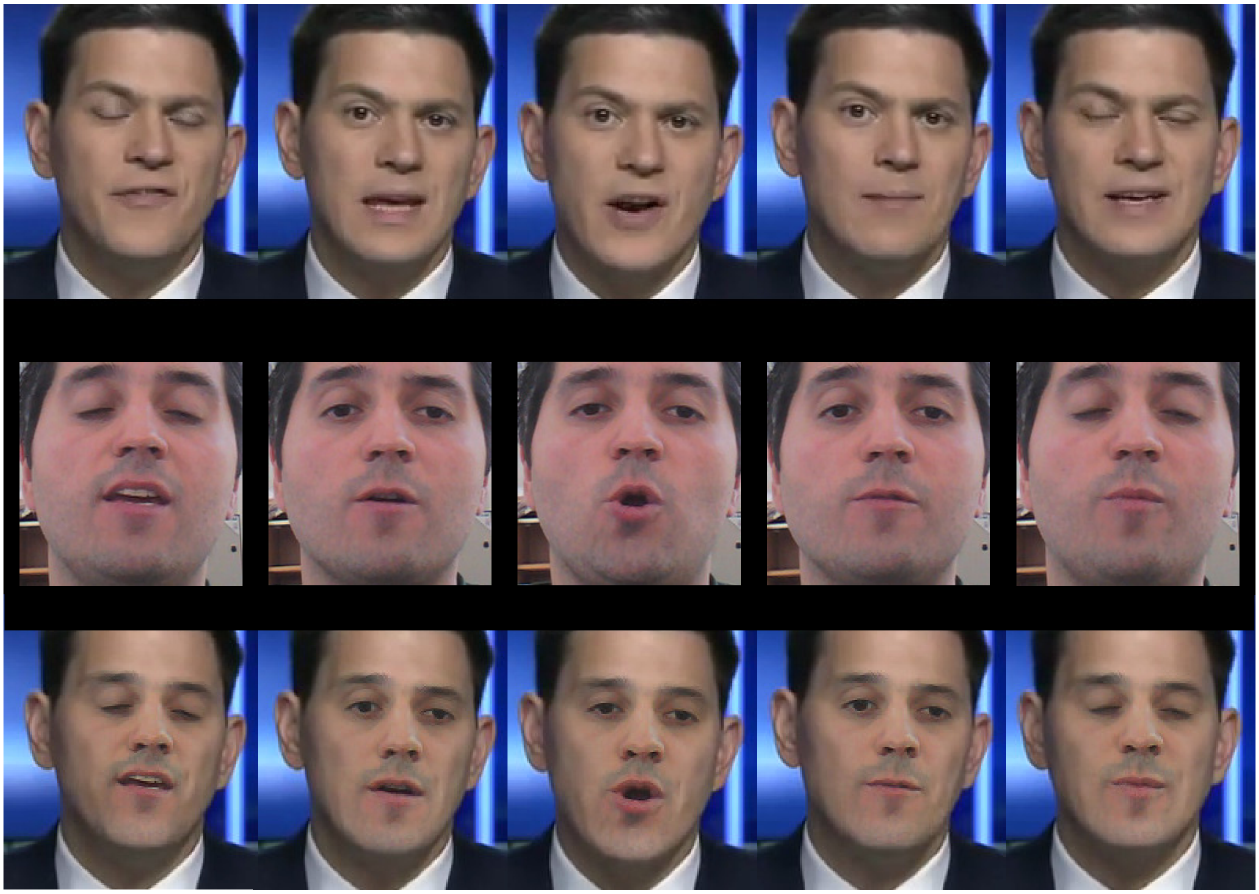

Figure 4. Low-quality video from the Internet (13 s of target footage, 10 s of source footage). David Miliband interview (http://youtu.be/uePg_ha7_jg). Top: Frames from the target sequence. Middle: Corresponding selected source frames. Bottom: Final composites. Chosen weights in Eq. (9) in the main paper: $w_{\text{nr}} = 0.55$, $w_{\text{r}} = 0.45$.

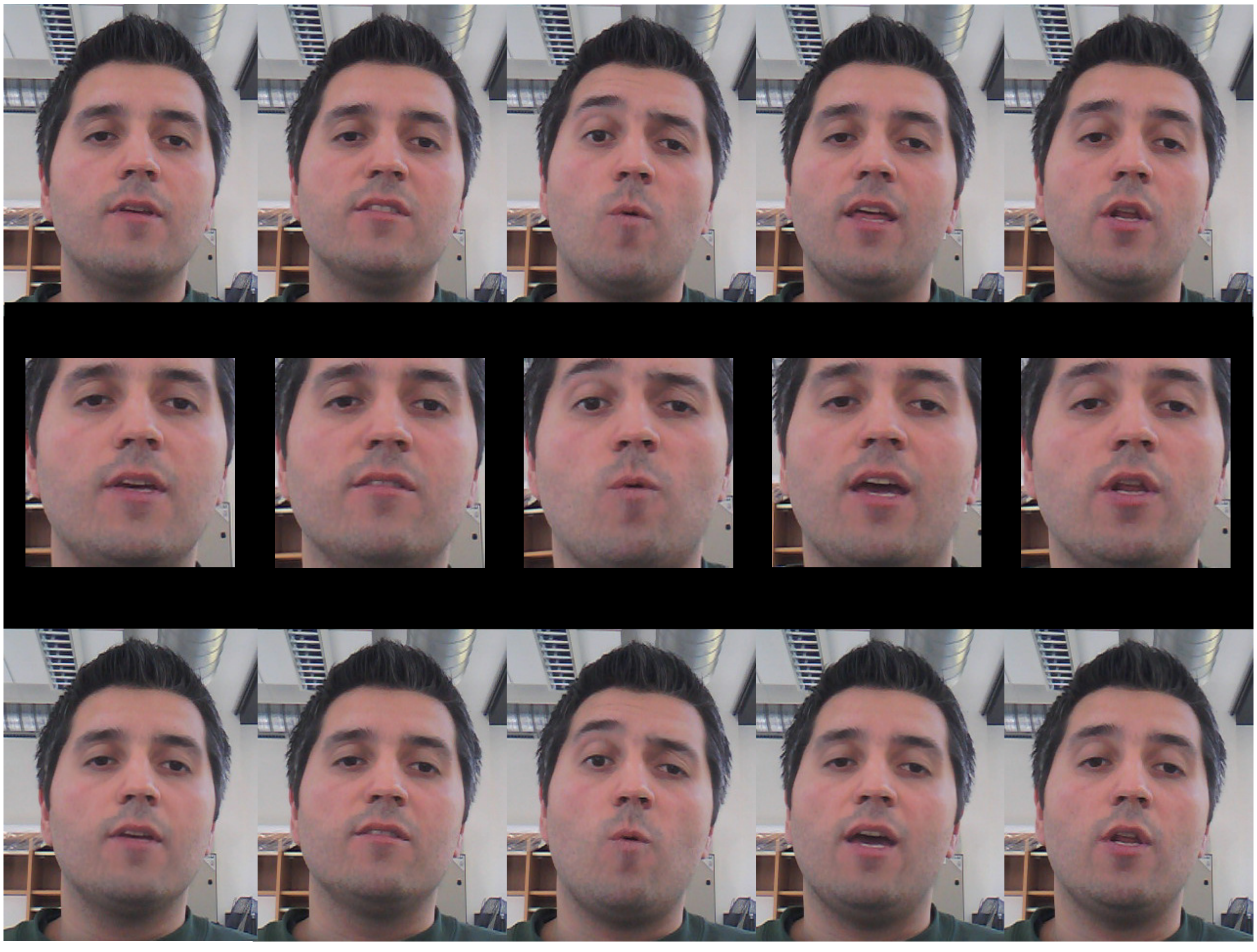

Figure 5. Self-reenactment of low-quality webcam video (10 s of target and source footage). Top: Frames from the target sequence. Middle: Corresponding selected source frames. Bottom: Final composites. Chosen weights in Eq. (9) in the main paper: $w_{\text{nr}} = 0.55$, $w_{\text{r}} = 0.45$.

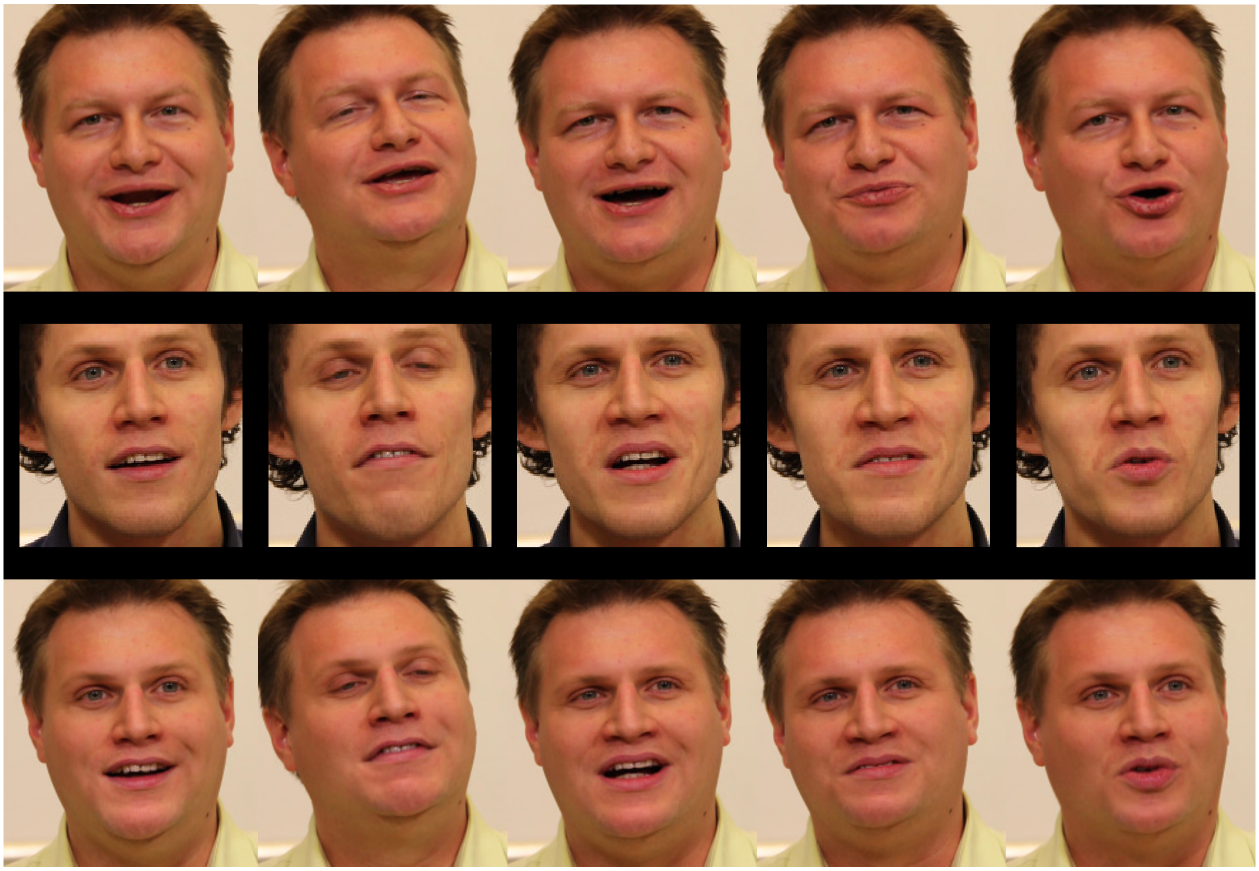

Figure 6. Existing high-quality video (17 s of target footage, 10 s of source footage). Top: Example frames from the target sequence. Middle: Corresponding selected source frames. Bottom: Final composites. Chosen weights in Eq. (9) in the main paper: $w_{\text{nr}} = 0.65$, $w_{\text{r}} = 0.35$.

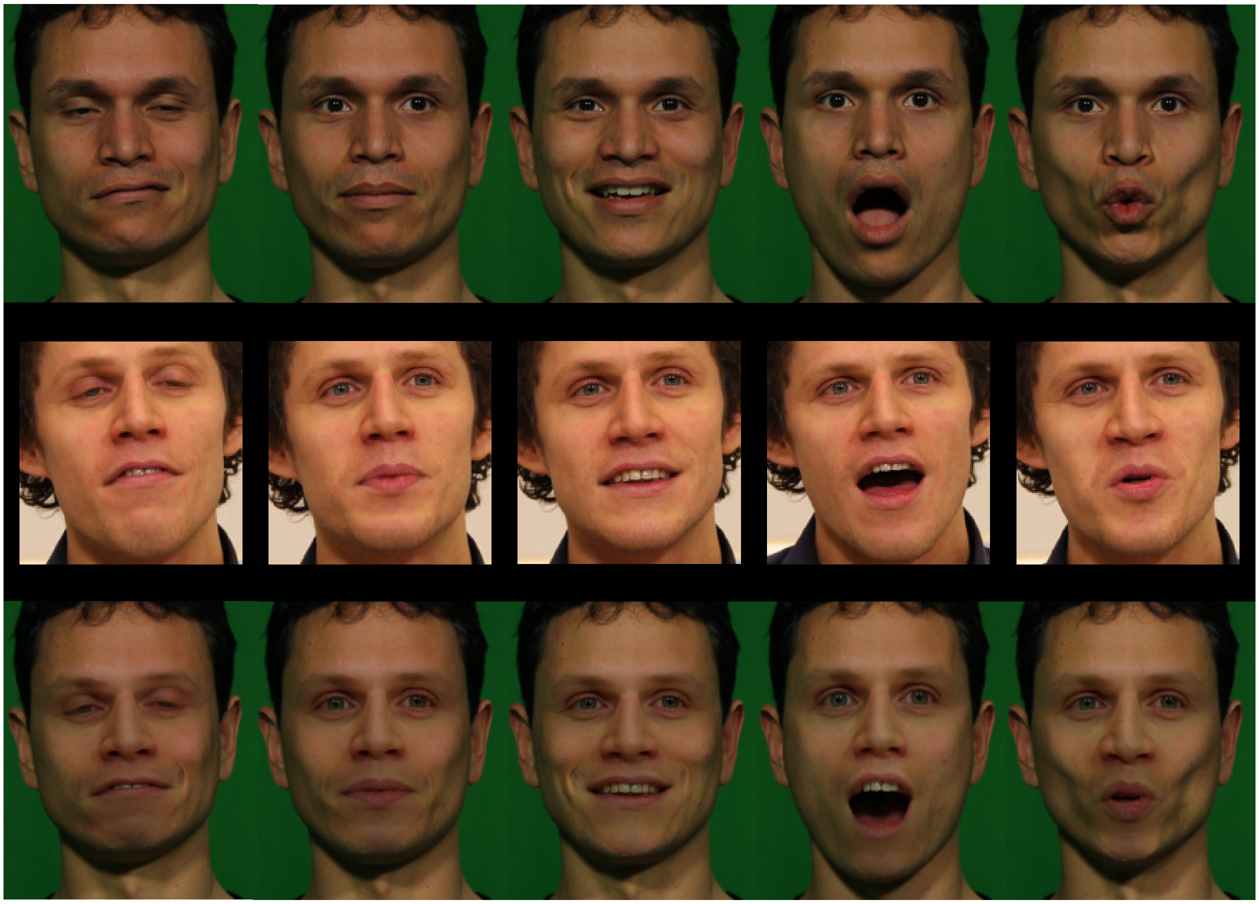

Figure 7. Existing high-quality video (22 s of target footage, 10 s of source footage). Top: Example frames from the target sequence. Middle: Corresponding selected source frames. Bottom: Final composites. Chosen weights in Eq. (9) in the main paper: $w_{\text{nr}} = 0.65$, $w_{\text{r}} = 0.35$.

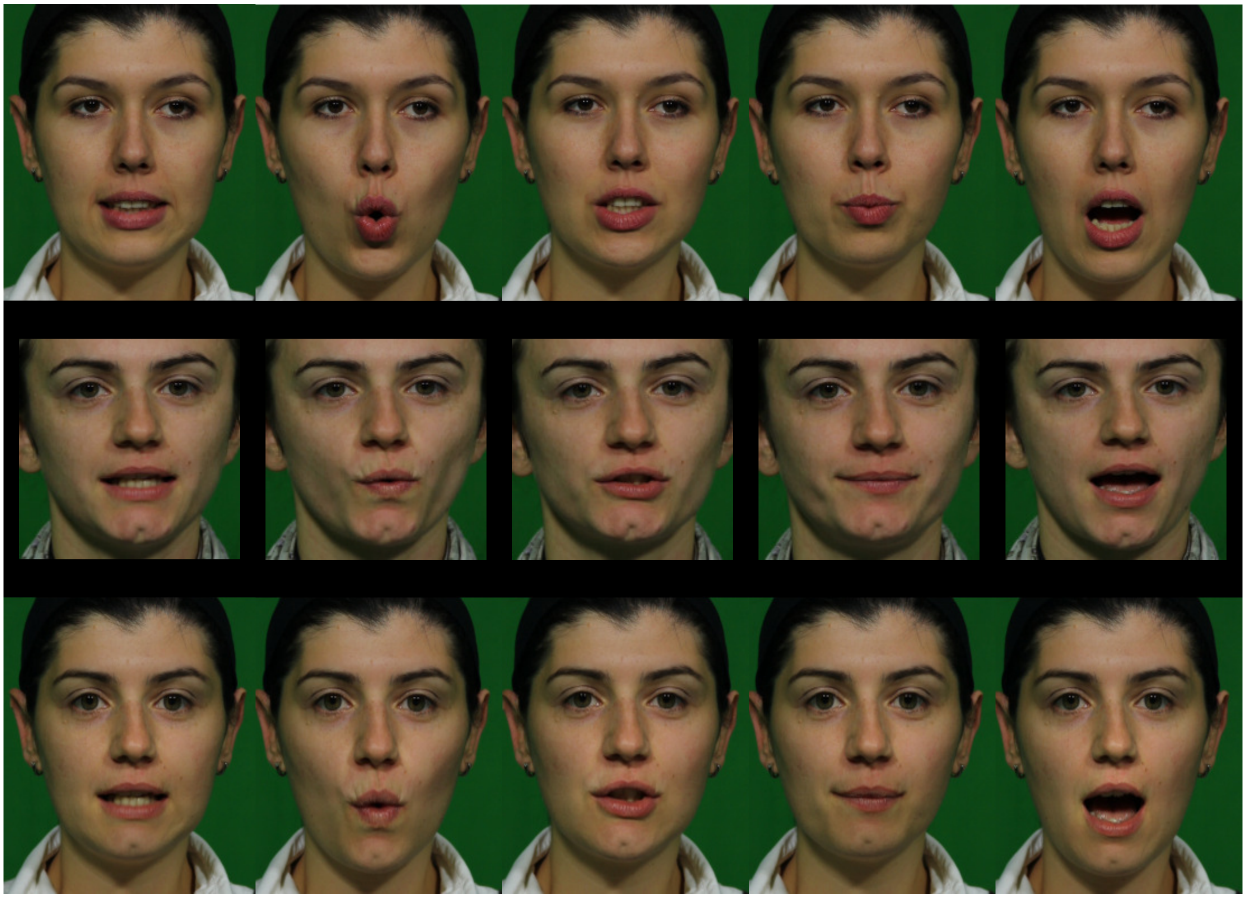

Figure 8. Existing high-quality video (14 s of target footage, 10 s of source footage). Top: Example frames from the target sequence. Middle: Corresponding selected source frames. Bottom: Final composites. Chosen weights in Eq. (9) in the main paper: $w_{\text{nr}} = 0.55$, $w_{\text{r}} = 0.45$.

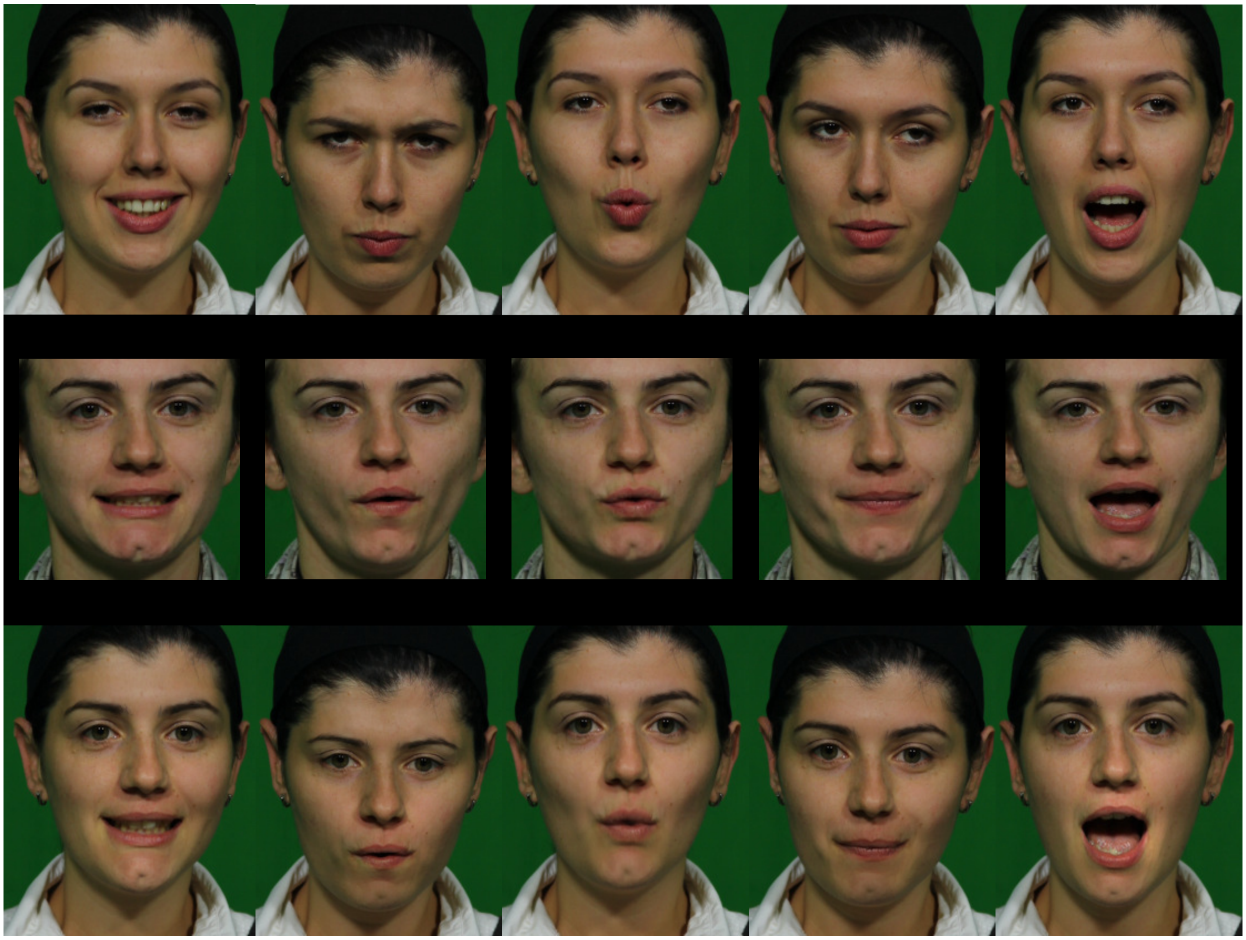

Figure 9. Existing high-quality video (12 s of target footage, 10 s of source footage). Top: Example frames from the target sequence. Middle: Corresponding selected source frames. Bottom: Final composites. Chosen weights in Eq. (9) in the main paper: $w_{\text{nr}} = 0.55$, $w_{\text{r}} = 0.45$.

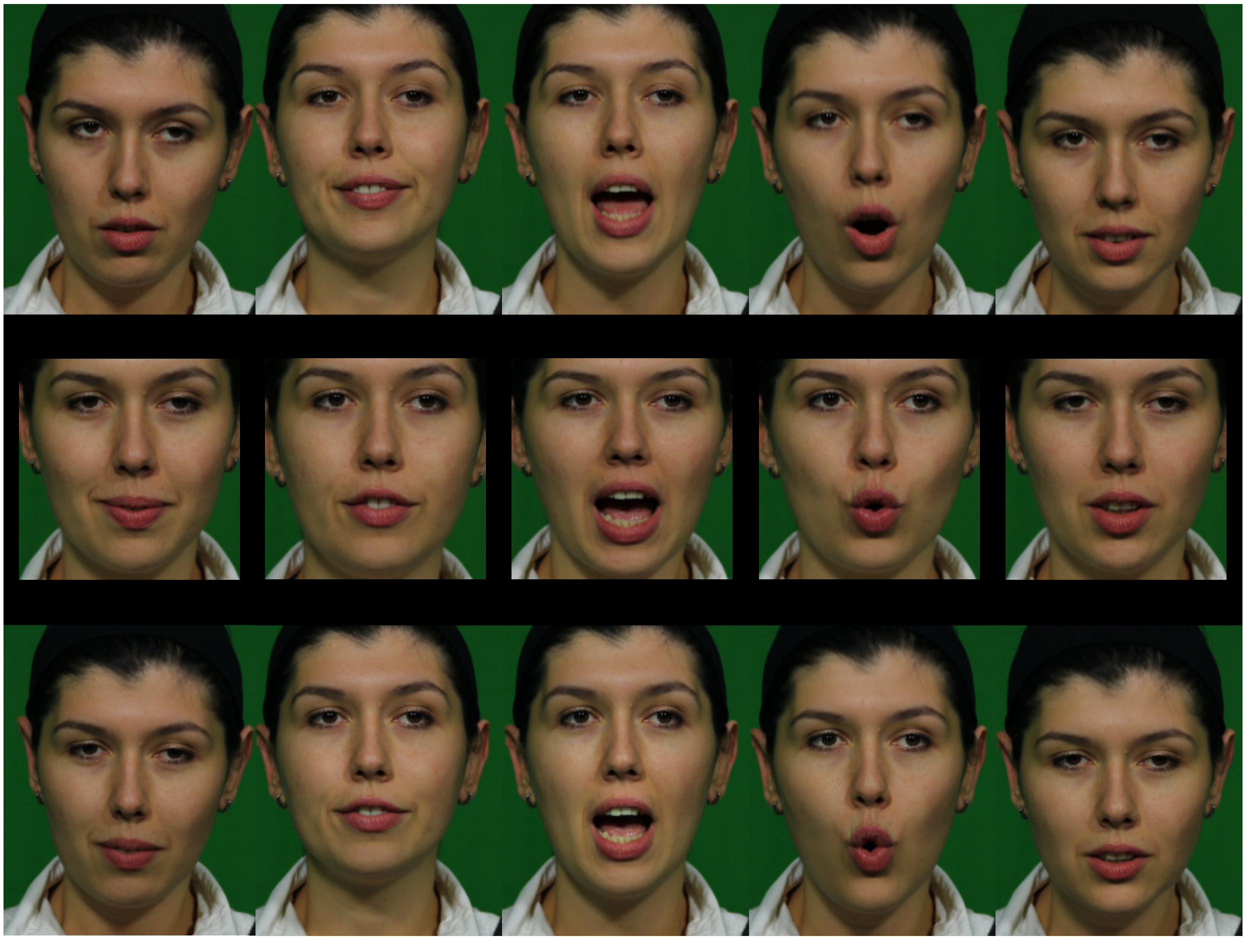

Figure 10. Reenactment of the same person under similar conditions in existing high-quality video (12 s of target footage, 14 s of source footage). Top: Example frames from the target sequence. Middle: Corresponding selected source frames. Bottom: Final composites. Chosen weights in Eq. (9) in the main paper: $w_{\text{nr}} = 0.55$, $w_{\text{r}} = 0.45$.

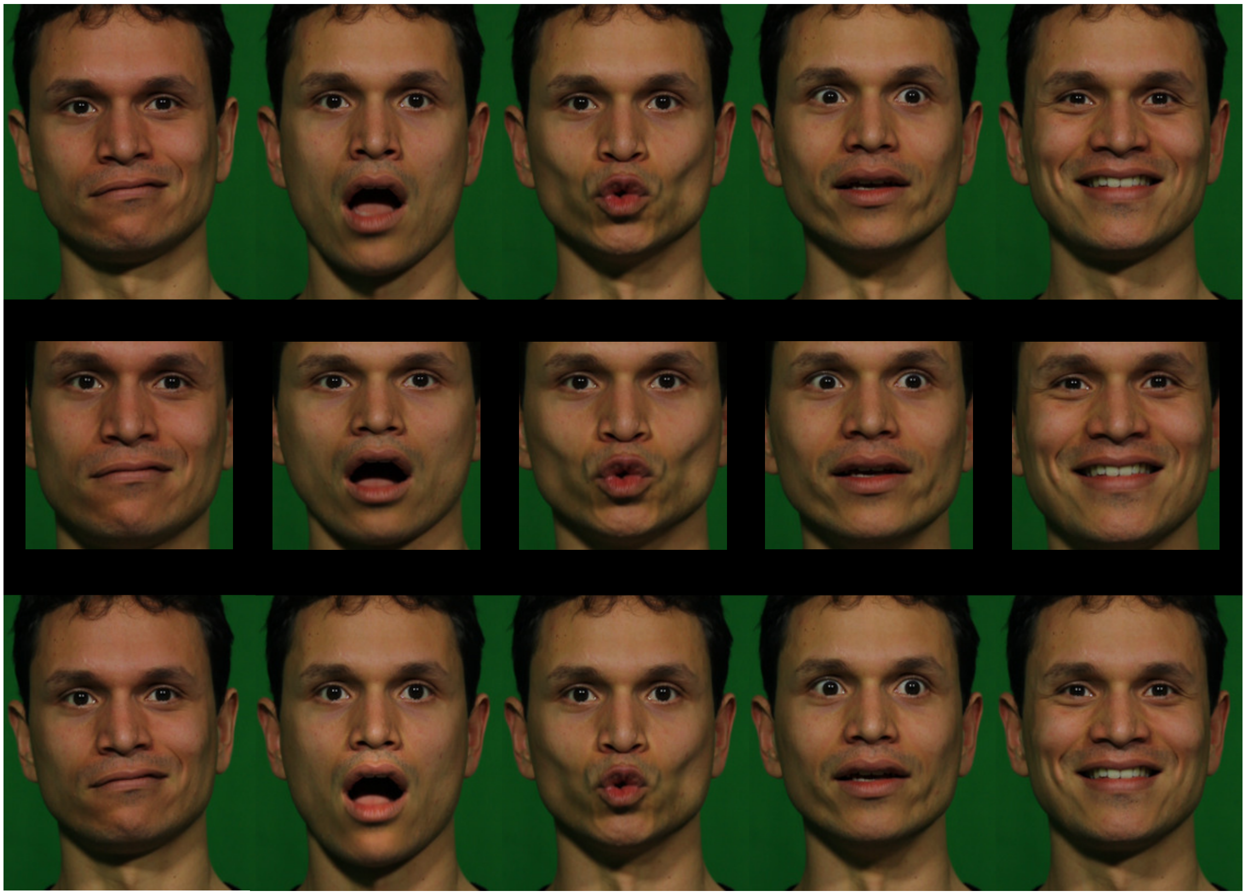

Figure 11. Self-reenactment of existing high-quality video (22 s of target and source footage). Top: Example frames from the target sequence. Middle: Corresponding selected source frames. Bottom: Final composites. Chosen weights in Eq. (9) in the main paper: $w_{\text{nr}} = 0.55$, $w_{\text{r}} = 0.45$.

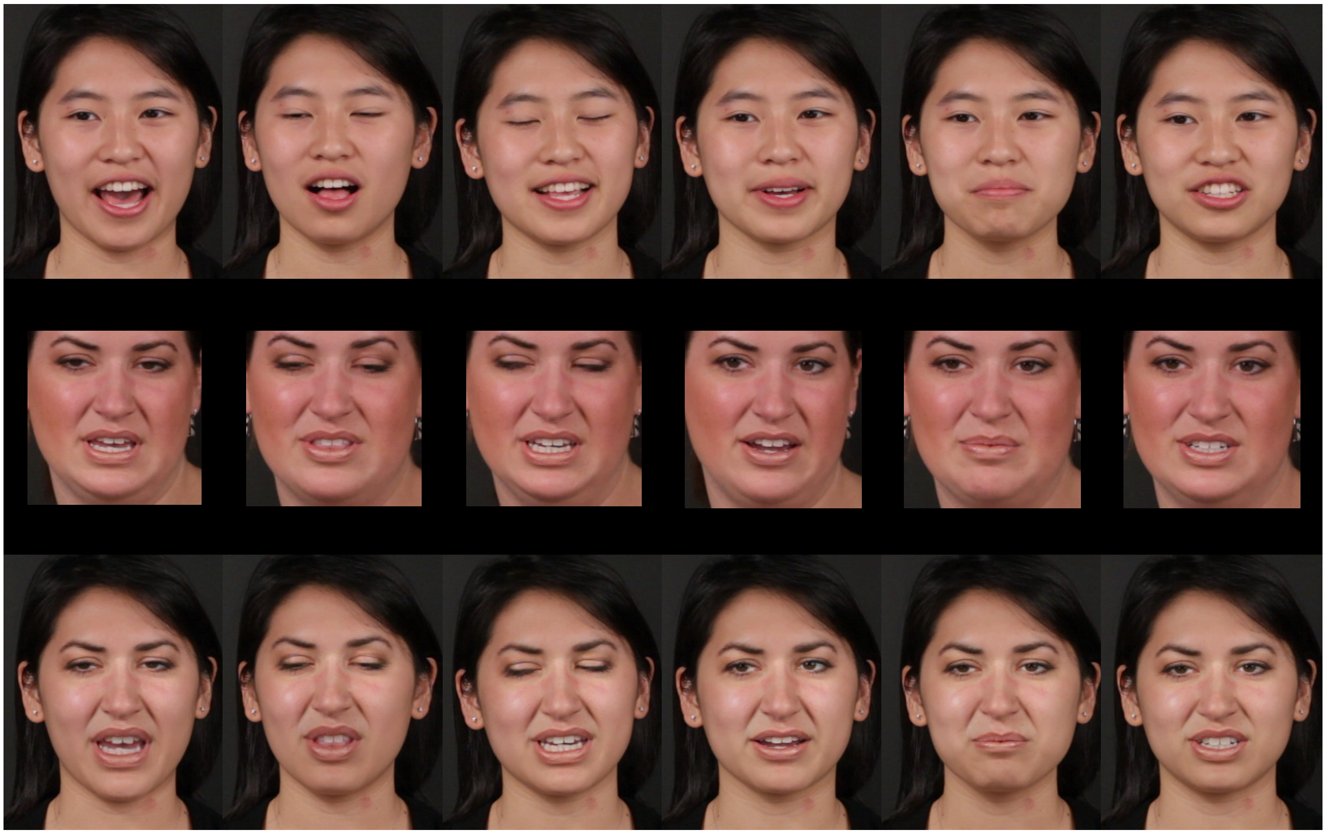

Figure 12. Comparison to Dale *et al*. [1] on existing high-quality video provided by the authors (17 s of target footage, 15 s of source footage). Top: Example frames from the target sequence. Middle: Corresponding selected source frames. Bottom: Final composites. Chosen weights in Eq. (9) in the main paper: $w_{\text{nr}} = 0.65$, $w_{\text{r}} = 0.35$.