Video Face Replacement

Kevin Dale[1] Kalyan Sunkavalli[1] Micah K. Johnson[2] Daniel Vlasic[3] Wojciech Matusik[2,4] Hanspeter Pfister[1]
[1]Harvard University [2]MIT CSAIL [3]Lantos Technologies [4]Disney Research Zurich

(a) Source (b) Target (c) Aligned (d) Three frames of the blended result

Figure 1: *Our method for face replacement requires only single-camera video of the source (a) and target (b) subject, which allows for simple acquisition and reuse of existing footage. We track both performances with a multilinear morphable model then spatially and temporally align the source face to the target footage (c). We then compute an optimal seam for gradient domain compositing that minimizes bleeding and flickering in the final result (d).*

Abstract

We present a method for replacing facial performances in video. Our approach accounts for differences in identity, visual appearance, speech, and timing between source and target videos. Unlike prior work, it does not require substantial manual operation or complex acquisition hardware, only single-camera video. We use a 3D multilinear model to track the facial performance in both videos. Using the corresponding 3D geometry, we warp the source to the target face and retime the source to match the target performance. We then compute an optimal seam through the video volume that maintains temporal consistency in the final composite. We showcase the use of our method on a variety of examples and present the result of a user study that suggests our results are difficult to distinguish from real video footage.

CR Categories: I.4.3 [Image Processing and Computer Vision]: Enhancement—Filtering; I.3.7 [Computer Graphics]: Three-Dimensional Graphics and Realism—Animation

Keywords: face replacement, facial animation, video compositing

1 Introduction

Techniques for manipulating and replacing faces in photographs have matured to the point that realistic results can be obtained with minimal user input (e.g., [Agarwala et al. 2004; Bitouk et al. 2008; Sunkavalli et al. 2010]). Face replacement in video, however, poses significant challenges due to the complex facial geometry as well as our perceptual sensitivity to both the static and dynamic elements of faces. As a result, current systems require complex hardware and significant user intervention to achieve a sufficient level of realism (e.g., [Alexander et al. 2009]).

This paper presents a method for face replacement in video that achieves high-quality results using a simple acquisition process. Unlike previous work, our approach assumes inexpensive hardware and requires minimal user intervention. Using a single camera and simple illumination, we capture *source* video that will be inserted into a *target* video (Fig. 1). We track the face in both the source and target videos using a 3D multilinear model. Then we warp the source video in both space and time to align it to the target. Finally, we blend the videos by computing an optimal spatio-temporal seam and a novel mesh-centric gradient domain blending technique.

Our system replaces all or part of the face in the target video with that from the source video. Source and target can have the same person or two different subjects. They can contain similar performances or two very different performances. And either the source or the target can be existing (i.e., uncontrolled) footage, as long as the face poses (i.e., rotation and translation) are approximately the same. This leads to a handful of unique and useful scenarios in film and video editing where video face replacement can be applied.

For example, it is common for multiple takes of the same scene to be shot in close succession during a television or movie shoot. While the timing of performances across takes is very similar, subtle variations in the actor's inflection or expression distinguish one take from the other. Instead of choosing the single best take for the final cut, our system can combine, e.g., the mouth performance from one take and the eyes, brow, and expressions from another to produce a *video montage*.

A related scenario is *dubbing*, where the source and target subject are the same, and the source video depicts an actor in a studio recording a foreign language track for the target footage shot on location. The resulting video face replacement can be far superior to the common approach of replacing the audio track only. In contrast to multi-take video montage, the timing of the dubbing source is completely different and the target face is typically fully replaced, although partial replacement of just the mouth performance is possible, too.

Another useful scenario involves *retargeting* existing footage to produce a sequence that combines an existing backdrop with a new face or places an existing actor's facial performance into new footage. Here the new footage is shot using the old footage as an audiovisual guide such that the timing of the performances roughly matches. Our video-based method is particularly suitable in this case because we have no control over the capture of the existing footage.

A final scenario is *replacement*, where the target facial performance is replaced with an arbitrary source performance by a different subject. This is useful, for example, when replacing a stunt actor's face, captured in a dangerous environment, with the star actor's face, recorded in a safe studio setting. In contrast to retargeting, where the source footage is shot using the target as an audiovisual guide to roughly match the timings, the performance of the source and target can be very different, similar to dubbing but with different subjects.

Furthermore, it is entertaining for amateurs to put faces of friends and family into popular movies or music videos. Indeed, an active community of users on YouTube has formed to share such videos despite the current manual process of creating them (e.g., search for "Obama Dance Off"). Our video face replacement system would certainly benefit these users by dramatically simplifying the currently labor-intensive process of making these videos.

Video face replacement has advantages over replacing the entire body or the head in video. Full body replacement typically requires chroma key compositing (i.e., green screening) or rotoscoping to separate the body from the video. Head replacement is difficult due to the complexities of determining an appropriate matte in regions containing hair. Existing methods for both body and head replacement require expensive equipment, significant manual work, or both [Alexander et al. 2009]. Such methods are not practical in an amateur setting and are also time consuming and challenging for professionals.

Our system does rely on a few assumptions about the input videos. It works best when the illumination in the source and target videos is similar. However, we mitigate this limitation by finding a coherent spatio-temporal seam for blending that minimizes the differences between the source and target videos (Sec. 6). Second, we assume that the pose of faces in the source and target videos is $\pm 45^o$ from frontal, otherwise automatic tracking and alignment of the faces will fail (Sec. 4). This assumption could be waived by employing user assistance during tracking.

The main contribution of this paper is a new system for video face replacement that does not require expensive equipment or significant user intervention. We developed a novel spatio-temporal seam finding technique that works on meshes for optimal coherent blending results. We demonstrate the applicability of our approach on a number of examples in four scenarios: video montage (Fig. 6), dubbing (Fig. 7), retargeting (Figs. 1 and 10), and replacement (Fig. 9). We present results of a user study on Mechanical Turk that demonstrates that our system is sufficient for plausible face replacement and difficult to distinguish from real footage (Sec. 7).

2 Previous Work

Face replacement in images and video has been considered in a variety of scenarios, including animation, expression transfer, and online privacy. However, the direct video-to-video face transfer presented in this paper has been relatively unexplored. We briefly describe previous work on face replacement and compare these approaches to our system.

Editing Faces in Images Face editing and replacement in images has been a subject of an extensive research. For example, the method by Blanz et al. [2004] fits a morphable model to faces in both the source and target images and renders the source face with the parameters estimated from the target image. The well-known photomontage [Agarwala et al. 2004] and instant cloning systems [Farbman et al. 2009] allow for replacing faces in photographs using seamless blending [Pérez et al. 2003]. Bitouk et al. [2008] describe a system for automatic face swapping using a large database of faces. They use this system to conceal the identity of the face in the target image. Face images have been also used as priors to enhance face attractiveness using global face warping [Leyvand et al. 2008] or to adjust tone, sharpness, and lighting of faces [Joshi et al. 2010]. The system of Sunkavalli et al. [2010] models the texture, noise, contrast and blur of the target face to improve the appearance of the composite. More recently, Yang et al. [2011] use optical flow to replace face expressions between two photographs. The flow is derived from 3D morphable models that are fit to the source and target photos. It is not clear whether any of these methods could achieve temporally coherent results when applied to a video sequence.

Face Replacement in Video using 3D Models The traditional way to replace faces in video is to acquire a 3D face model of the actor, to animate the face, and to relight, render, and composite the animated model into the source footage. The 3D face model of the actor can be captured using marker-based [Williams 1990; Guenter et al. 1998; Bickel et al. 2007], structured light [Zhang et al. 2004; Ma et al. 2008; Li et al. 2009; Weise et al. 2009], or passive multi-view stereo approaches [Jones et al. 2006; Bradley et al. 2010; Beeler et al. 2011 (to appear)]. Model-based face replacement can achieve remarkable realism. Notable examples include the recreation of actors for The Matrix Reloaded [Borshukov et al. 2003], The Curious Case of Benjamin Button [Robertson 2009], and the Digital Emily project [Alexander et al. 2009]. However, these methods are expensive, and typically require complex hardware and significant user intervention to achieve a sufficient level of realism.

Video-to-Video Face Replacement Purely image-based methods do not construct a 3D model of the actor. Bregler et al. [1997] and Ezzat et al. [2002] replace the mouth region in video to match phonemes of novel audio input using a database of training images of the same actor. Flagg et al. [2009] use video-textures to synthesize plausible articulated body motion. Kemelmacher-Shlizerman et al. [2010] make use of image collections and videos of celebrities available online and replace face photos in real-time based on expression and pose similarity. However, none of these methods are able to synthesize the subtleties of the facial performance of an actor.

Morphable-Models for Face Synthesis Closely related to our work are image-based face capture methods [Essa et al. 1996; DeCarlo and Metaxas 1996; Pighin et al. 1999; Blanz et al. 2003; Vlasic et al. 2005]. These approaches build a morphable 3D face model from source images without markers or special face scanning equipment. We use the multilinear model by Vlasic at al. [2005] that captures identity, expression, and visemes in the source and target videos. Existing approaches use the estimated model parameters to generate and drive a detailed 3D textured face mesh for a target identity, which can be seamlessly rendered back into target footage. In general, these systems assume the source actor's performance, but not their face, is desired in the newly synthesized output video. In contrast, our approach blends the source actor's complete face and performance, with all of its nuances intact, into the target video.

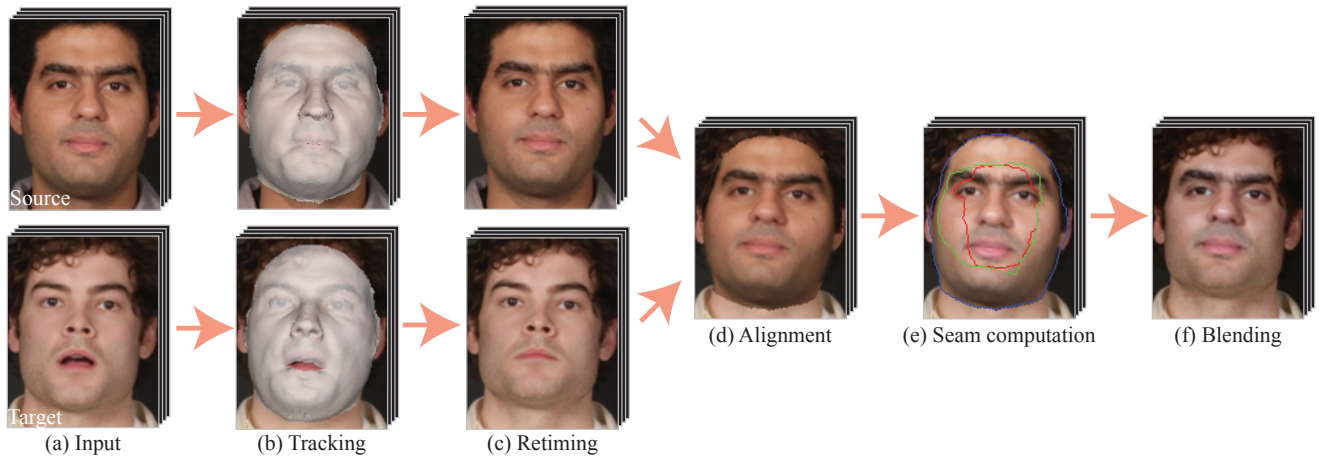


Figure 2: *An overview of our method. (a) Existing footage or single camera video serves as input source and target videos. (b) Both sequences are tracked and (c) optionally retimed to temporally align the performances. (d) The source face is spatially aligned in the target video. (e) An optimal seam is computed through the target video to minimize blending artifacts, and (f) the final composite is created with gradient-domain blending.*

3 Overview

Figure 2 shows an overview of our method. In order to replace a source face with a target face, we first model and track facial performances of both source and target with the multilinear method and data of Vlasic et al. [2005]. Their method estimates a multilinear model from 3D face scans of different identities, expressions, and speech articulations (i.e., visemes). It tracks parameters for these attributes and the 3D pose of the face (given as a rotation, translation, and scale) over a video sequence. At each frame, the pose, the multilinear model, and its parameters can be used to generate a 3D mesh that matches the geometry of the subject's face. A sufficient approximate fit is obtainable even for new faces that are not present in the original dataset. We reprocessed the original training data from Vlasic et al. covering 16 identities $\times$ 5 expressions $\times$ 5 visemes–a total of 400 face scans–placing them into correspondence with a face mesh that extends beyond the jaw and chin regions (Sec. 7).

In some scenarios it is important that the timing of the facial performance matches precisely in the source and the target footage. However, it might be very tedious to match these timings exactly as demonstrated by the numerous takes that are typically necessary to obtain compelling voiceovers (e.g., when re-recording a dialog for a film.) Instead, we only require a coarse synchronization between source and target videos and automatically retime the footage to generate a precise match for the replacement.

After tracking and retiming, we blend the source performance into the target video to produce the final result. This blending makes use of gradient-domain compositing to merge the source actor's face into the target video. While gradient domain compositing can produce realistic seamless results, the quality of the composite is often tied to the seam along which the blend is computed. Using an arbitrary seam is known to lead to bleeding artifacts. To minimize these artifacts we automatically compute an optimal spatio-temporal seam through the source and target that minimizes the difference across the seam on the face mesh and ensure that the regions being combined are compatible. In the second stage we use this seam to merge the gradients and recover the final composite video. For the results shown in the paper, each of which is about 10 seconds, processing requires about 20 minutes.

4 Face Tracking

Input Footage for all examples, except those that reuse existing footage, was captured with a Canon T2i camera with 85 mm and 50 mm lenses at 30 frames per second. In-lab sequences were lit with 300 W studio lights placed on the left and right and in front of the subject, softened by umbrella reflectors. When appropriate, we used the target video as an audio-visual guide during capture of the source (or vice versa) to approximately match timing. All such examples in this paper were captured in 1-2 takes. For pose, actors were simply instructed to face the camera; natural head motion is accounted for with tracking.

Tracking To track a face across a sequence of frames, the method of Vlasic et al. [2005] computes the pose and attribute parameters of the multilinear face model that best explain the optical flow between adjacent frames in the sequence. The multilinear face model $\mathscr{M}$, an N-mode tensor with a total of $3K \times D_2 \times \ldots \times D_N$ elements (where K is the number of vertices in a single face mesh), is obtained via N-mode singular value decomposition (N-mode SVD) from the N-mode data tensor containing the vertex positions of the original scan data (the Cartesian product over expression, viseme, and identity).

With the multilinear model in hand, the original face data can be interpolated or extrapolated to generate a new face as

$$\mathbf{f} = \mathscr{M} \times_2 \mathbf{w}_2^\top \times_3 \mathbf{w}_3^\top \times_4 \mathbf{w}_4^\top, \qquad (1)$$

where mode 1 corresponds to vertex positions in the 4-mode model, $\mathbf{w}_i$ is a $D_i \times 1$ column vector of parameters for the attribute corresponding to the i^{th} mode (i.e., one of expression, viseme, or identity), $\mathbf{f}$ is a $3K$-element column vector of new vertex positions, and the $\times_n$ operator is the mode-n product, defined between a tensor and a matrix. We refer the reader to Vlasic et al. [2005] for more details.

Initialization Since tracking is based on optical flow, initialization is critical, as errors in the initialization will be propagated throughout the sequence. Moreover, tracking can go astray on troublesome frames, e.g., due to motion blur, extreme pose change, high frequency lighting, or occlusions. Therefore, we also provide a simple

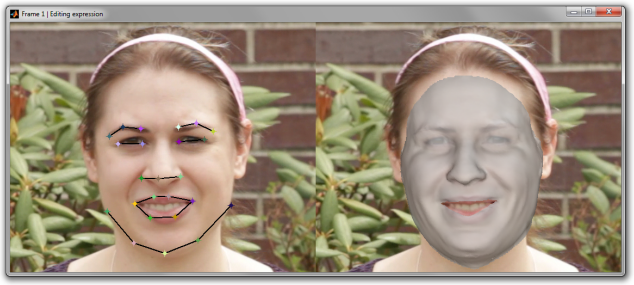


Figure 3: *User interface for tracking. To refine the initialization or correct tracking at a specific key frame, the user can adjust a few markers on the face to adjust pose, expression, or viseme.*

user interface that can ensure good initialization and can correct tracking for troublesome frames.

The interface allows the user to adjust positions of markers on the eyes, eyebrows, nose, mouth, and jawline, from which the best-fit pose and model parameters are computed. The user can alternate between adjusting pose and each attribute individually; typically, 1 iteration of each is sufficient for good initialization (Fig. 3).

We start by automatically detecting the face [Viola and Jones 2001]. Next, we localize facial features [Everingham et al. 2006] (e.g., the corners of the mouth, eyes, and nose) in the first frame of a sequence. Then, we compute the initial pose that best aligns the detected features with the corresponding source features in the face mesh. This initial face mesh is generated from the multilinear model using a user-specified set of initial attributes corresponding to the most appropriate expression, viseme, and identity.

Holding all but one attribute's parameters fixed, we can project the multilinear model $\mathscr{M}$ onto the subspace corresponding to the remaining attribute, e.g., for the third attribute:

$$\mathbf{A}_3 = \mathscr{M} \times_2 \mathbf{w}_2^\top \times_4 \mathbf{w}_4^\top, \tag{2}$$

for the $3K \times D_3$ matrix $\mathbf{A}_3$. Given $\mathbf{A}_i$ and a column vector $\mathbf{g}$ of target vertex positions, we can compute parameters for the i^{th} attribute that best fit the target geometry as

$$\underset{\mathbf{w}_i}{\operatorname{argmin}} \|\mathbf{g} - \mathbf{A}_i \mathbf{w}_i\|^2. \tag{3}$$

The least squares solution to Eq. 3 is given as

$$\mathbf{w}_i = (\mathbf{A}_i^\top \mathbf{A}_i)^{-1} \mathbf{A}_i^\top \mathbf{g}. \tag{4}$$

To fit parameters for the i^{th} attribute to image space markers, we take the subset of the multilinear model corresponding to the (x, y) coordinates of mesh vertices that should align to the markers and apply Eq. 4, populating $\mathbf{g}$ with marker positions, transformed to the coordinate frame of the model via an inverse pose transformation.

While multilinear tracking does well at tracking expression and viseme, which vary from frame to frame, we found that identity, which is computed over the full sequence and held constant, was not. Even after multiple iterations of tracking, each of which updates identity parameters, those parameters changed very little from their initial values. This caused significant problems when tracking with a full face model, where it is critical that the mesh covers the subject's entire face, and only their face (no background) over the entire sequence. Therefore it is important to have an accurate initialization of identity.

We employ the FaceGen Modeller [Singular Inversions Inc. 2011] in order to obtain a better initialization of the identity parameters. FaceGen generates a 3D mesh based on a frontal face image and, optionally, a profile image. The input images can be extracted from the original video sequences or downloaded from the Internet when reusing existing footage. The input images need to depict the subject with a closed-mouth neutral expression. FaceGen requires minimal user input to specify about 10 markers per image. All meshes created by FaceGen are themselves in correspondence. Therefore, we can register the FaceGen mesh with the multilinear model using the same template-fitting procedure [Vlasic et al. 2005] we used to register the original scan data. We then fit the multilinear model to the registered FaceGen mesh using procrustes alignments to our current best-fit mesh and using Eqs. 3 and 4 to solve for the best-fit identity parameters. In this optimization we only use about 1 percent of the original mesh vertices. The process typically converges in 10 iterations.

Key framing We can use the same interface (Fig. 3) for adjusting pose and attribute parameters at specific key frames where automatic tracking fails. First, we track the subsequences between each pair of user-adjusted key frames in both the forward and reverse directions and linearly interpolate the two results. We then perform additional tracking iterations on the full sequence to refine pose and parameter estimates across key frame boundaries. Note that none of the results shown in the paper required key framing.

5 Spatial and Temporal Alignment

Spatial alignment From an image sequence I, where $I(x, t)$ denotes the value at pixel position x in frame t, tracking produces a sequence of attribute parameters and pose transformations. For each frame t, $\mathbf{f}(t)$ is the column vector of vertex positions computed from attribute parameters at time t using Eq. 1, and $\mathbf{f}_i(t)$, the i^{th} vertex at time t. Per-frame pose consists of a scale s, 3×3 rotation matrix $\mathbf{R}$, and a translation vector $\mathbf{t}$ that together transform the face meshes into their tracked positions in image space coordinates. Subscripts S and T denote source and target, respectively.

To align the source face in the target frame, we use the face geometry from the source sequence and pose from the target sequence. That is, for frame t, the aligned position of the i^{th} source vertex position is given as

$$\mathbf{f'}_{i,S}(t) = s_T(t)\mathbf{R}_T(t)\mathbf{f}_{i,S}(t) + \mathbf{t}_T(t) \tag{5}$$

We also take texture from the source image I_S; texture coordinates are computed similarly to Eq. 5 using instead both source geometry and source pose.

While we track the full face mesh in both source and target sequences, the user may choose to replace only part of the target face, for example, in the multi-take video montage result in Fig. 6. In this case, the user either selects from a predefined set of masks – eyes, eyes and nose, or mouth – or paints an arbitrary mask on the face. In these cases, $\mathbf{f}'_S$ represents only those vertices within the user-specified mask.

Retiming We retime the footage using Dynamic Time Warping (DTW) [Rabiner and Juang 1993]. DTW is a dynamic programming algorithm that seeks a monotonic mapping between two sequences that minimizes the total cost of pairwise mappings. The output of DTW provides a reordering of one sequence to best match the other. Here we define pairwise cost between source and target frames according to the motion of the mouth in each frame. We found that computing cost based on motion instead of absolute position was more robust across differences in mouth shape and articulation in different subjects.

Specifically, for the loop of vertices along the interior of the upper and lower lip, we compare the average minimum Euclidean distance between the first partial derivatives with respect to time. Comparing velocity of mouth vertices for this step, as opposed to position, ensures robustness to differences in mouth shape between source and target. We compute these partial derivatives using first order differencing on the original vertex positions without transforming to image space. Let $\mathbf{m}_{i,S}(t_1)$ and $\mathbf{m}_{j,T}(t_2)$ be the partial derivatives for the i^{th} vertex in the source mouth at time t_1 and the j^{th} vertex in the target mouth at time t_2 , respectively. Then the cost of mapping source frame t_1 to target frame t_2 for DTW is

$$\sum_i \min_j ||\mathbf{m}_{i,S}(t_1) - \mathbf{m}_{j,T}(t_2)|| + \min_j ||\mathbf{m}_{j,S}(t_1) - \mathbf{m}_{i,T}(t_2)||. \tag{6}$$

DTW does not consider temporal continuity. The resulting mapping may include 'stairstepping', where a given frame is repeated multiple times, followed by a non-consecutive frame, which appears unnatural in the retimed video. We smooth the mapping with a low-pass filter and round the result to the nearest integer frame. This maintains sufficient synchronization while removing discontinuities. While there are more sophisticated methods that can directly enforce continuity e.g., Hidden Markov Models (HMMs), as well as those for temporal resampling, we found this approach to be fast and well-suited to our input data, where timing is already fairly close.

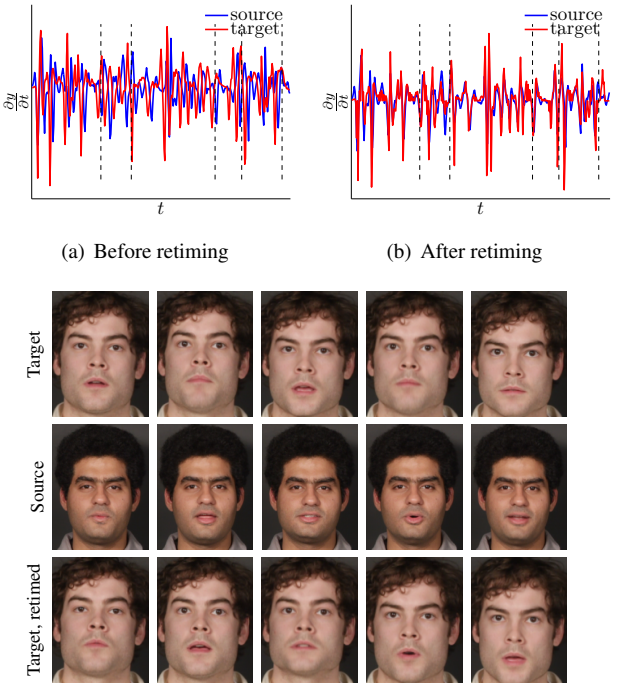
Since the timing of the original source and target videos is already close, the mapping can be applied from source to target and vice versa (for example, to maintain important motion in the background of the target or to capture the subtle timing of the source actor's performance.) For simplicity, in the following sections $\mathbf{f}_S(t)$ and $\mathbf{f}_T(t)$, as well as their corresponding texture coordinates and texture data, refer to the retimed sequences when retiming is employed and to the original sequences when it is not. Fig. 4 highlights the result of retiming inputs with dialog with DTW.

6 Blending

Optimal Seam Finding Having aligned the source face texture to the target face, we would like to create a truly photo-realistic composite by blending the two together. While this can be accomplishing using gradient-domain fusion [Pérez et al. 2003], we need to specify the region from the aligned video that needs to be blended into the target video, or alternatively, the *seam* that demarcates the region in the composite that comes from the target video from the region that comes from the aligned video. While the edge of face mesh could be used as the seam, in many cases it cuts across features in the video leading to artifacts such as bleeding (see Fig. 5). In addition, this seam needs to be specified in every frame of the composite video, making it very tedious for the user to do.

We solve this problem by automatically estimating a seam in space-time that minimizes the differences between the aligned and target videos, thereby avoiding bleeding artifacts. While a similar issue has been addressed in previous work [Jia et al. 2006; Agarwala et al. 2004; Kwatra et al. 2003], our problem has two important differences. First, the faces we are blending often undergo large (rigid and non-rigid) transformations, and the seam computation needs to be handle this. Second, it is important that the seam be temporally coherent to ensure that the composited region does not change substantially from frame to frame leading to flickering artifacts (see Fig. 5).

Our algorithm incorporates these requirements in a novel graph-cut framework that estimates the optimal seam *on the face mesh*. For every frame in the video, we compute a closed polygon on the face mesh that separates the source region from the target region; projecting this polygon onto the frame gives us ths corresponding *image-space* seam. Estimating the seam in *mesh-space* helps us handle our two requirements. First, when the face deforms in the source and target videos, the face mesh deforms to track it without any changes in its topology. The mesh already accounts for these deformations, making the seam computation invariant to these changes. For example, when a subject talks, the vertices corresponding to his lips remain the same, while their positions change. Thus, a polygon corresponding to these vertices defines a time-varying seam that stays true to the motion of the mouth. Second, estimating the seam on the mesh allows us to enforce temporal constraints that encourage the seam to pass through the same vertices over time. Since the face vertices track the same face features over time this means that same parts of the face are preserved from the source video in every frame.



(a) Before retiming (b) After retiming

(c) Corresponding frames before and after retiming

Figure 4: *Motion of the center vertex of the lower lip for source and target before retiming (a) and after (b). Corresponding cropped frames before and after retiming (c).*

We formulate the optimal seam computation as a problem of labeling the vertices of the face mesh as belonging to the source or target video. We do this by constructing a graph on the basis of the face mesh and computing the min-cut of this graph. The nodes of this graph correspond to the vertices in the face aligned mesh over time (i.e., $\mathbf{f}_i(t) \forall i, t$). The edges in the graph consist of spatial edges corresponding to the edges in the mesh (i.e., all the edges between a vertex $\mathbf{f}_i(t)$ and its neighbor $\mathbf{f}_j(t)$) as well as temporal edges between corresponding vertices from frame to frame (i.e., between $\mathbf{f}_i(t)$ and $\mathbf{f}_i(t+1)$).

Similar to previous work on graphcut textures [Kwatra et al. 2003] and photomontage [Agarwala et al. 2004], we want the seam to cut through edges where the differences between the source and target video frames are minimal. This is done by setting the weights on the spatial edges in the graph between neighboring vertices $\mathbf{f}_i(t)$

Figure 5: *Seam computation for blending. The face mask boundary (blue), user-specified region to be preserved (red), and the optimal seam (green) are marked in each source frame. (a) Directly blending the source and target produces results with strong bleeding artifacts. (b) Computing a seam in image space improves results substantially but does not vary as pose and expression change. (c) A seam computed on the mesh can track these variations but may lead to flickering artifacts (see accompanying video) without additional constraints. (d) Enforcing temporal coherence minimizes these artifacts.*

and $\mathbf{f}_j(t)$ as:

$$W_s(\mathbf{f}_i(t), \mathbf{f}_j(t)) = ||I_S(\mathbf{f}_i(t), t) - I_T(\mathbf{f}_i(t), t)|| + ||I_S(\mathbf{f}_j(t), t) - I_T(\mathbf{f}_j(t), t)|| \quad (7)$$

When both the source and the target videos have very similar pixel values at vertices $\mathbf{f}_i(t)$ and $\mathbf{f}_j(t)$, the corresponding weight term takes on a very small value. This makes it favorable for a min-cut to cut across this this edge.

We would also like the seam to stay temporally coherent to ensure that the final composite does not flicker. We ensure this by setting the weights for the temporal edges of the graph as follows:

$$\begin{aligned} W_t(\mathbf{f}_i(t), \mathbf{f}_i(t+1)) &= W(\mathbf{f}_i(t+1), \mathbf{f}_i(t)) \quad (8) \\ &= \lambda(||I_S(\mathbf{f}_i(t), t) - I_S(\mathbf{f}_i(t), t+1)||^{-1} \\ &\quad + ||I_T(\mathbf{f}_i, t) - I_T\mathbf{f}_i, t+1)||^{-1}), \end{aligned}$$

where λ is used to control the influence of the temporal coherence. Unlike the spatial weights, these weights are constructed to have high values when the appearance of the vertices doesn't change much over time. If the appearance of vertex $\mathbf{f}_i(t)$ does not change over time in either the source or target video, this weight term takes on a large value, thus making it unlikely that the min-cut would pass through this edge, thus ensuring that this vertex has the same label over time. However, if the appearance of the vertex does change (due to the appearance of features such as hair, eyebrows, etc.), the temporal weight drops. This makes the seam temporally coherent while retaining the ability to shift to avoid features that cause large differences in intensity values. In practice, we set λ as the ratio of the sum of the spatial and temporal weights, i.e., $\lambda = \sum_{i,j,t} W_s(\mathbf{f}_i(t), \mathbf{f}_j(t), t) / \sum_{i,j,t} W_t(\mathbf{f}_i(t), \mathbf{f}_i(t+1))$. This ensures that the spatial and temporal terms are weighted approximately equally.

The vertices on the boundary of the face mesh in every frame are labeled as target vertices as they definitely come from the target videos. Similarly, a small set of vertices in the interior of the mesh are labeled as source vertices. This set can be directly specified by the user in one single frame.

Having constructed this graph, we use the alpha-expansion algorithm [Boykov et al. 2001] to label the mesh vertices as belonging to the either the source or target videos. The construction of the graph ensures that, in every frame, the graph-cut seam forms a closed polygon that separates the target vertices from the source vertices. From these labels we can explicitly compute this closed polygon $\partial P(t) = \{p_0(t), p_1(t), \cdots, p_{m_t}(t)\}$ for every frame. In addition, we also project these labels onto the frames to compute the corresponding image-space mask for compositing.

Fig. 5 shows the results of estimating the seam using our technique on an example video sequence. As can be seen in this example, using the edge of the face mesh as the seam leads to strong bleeding artifacts. Computing an optimal seam ensures that these artifacts don't occur. However, without temporal coherence, the optimal seam "jumps" from frame to frame, leading to flickering in the video. By computing the seam on the mesh using our combination of spatial and temporal weights we are able to produce a realistic composite that stays coherent over time. Please see the accompanying video to observe these effects.

Compositing Having estimated the optimal seam for compositing, we blend the source and target videos using gradient-domain fusion. We do this using a recently proposed technique that uses mean value coordinates [Farbman et al. 2009] to interpolate the differences between the source and target frames along the boundary. We re-use the face mesh to interpolate these differences. In particular, for every frame of the video, we compute the differences between source and target frames along the seam $\partial P(t)$, and interpolate them at the remaining source vertices using mean value coordinates. These differences are then projected onto the image and added to the source video to compute the final blended composite video.

7 Results and Discussion

Results We show results for a number of different subjects, capture conditions, and replacement scenarios. Fig. 6 shows multi-take video montage examples, both shot outdoors with a handheld camera. Fig. 7 shows dubbing results of a translation scenario, where the source and target depict the same subject speaking in different languages, with source captured in a studio setting and target captured outdoors. Figs. 9 shows a replacement result with different source and target subjects and notably different performances. Fig. 10 shows a retargeting result with different subjects, where the target was used as an audiovisual guide and the source retimed to match the target.

User interaction Although the majority of our system is automatic, some user interaction is required. This includes placing markers in FaceGen, adjusting markers for tracking initialization, and specifying the initial blending mask. Interaction in FaceGen required 2-3 minutes per subject. Tracking initialization was performed in less than a minute for all videos used in our results; the amount of interaction here depends on the accuracy of the automatic face detection and the degree to which the subject's expression and viseme differ from closed-mouth neutral. Finally, specifying the mask for blending in the first frame of every example took between 30 seconds and 1 minute. For any given result, total interaction time is therefore on the order of a few minutes, which is significantly less than what would be required using existing video compositing methods.

Comparison with Vlasic et al. [2005] We reprocessed the original scan data [Vlasic et al. 2005] to place it into correspondence with a face mesh that covers the full face, including the jaw. This was done for two reasons. First, the original model only covered the interior of the face; this restricted us to scenarios where the timing of the source and target's mouth motion must match exactly. While this is the case for multi-take montage and some dubbing scenarios when

Figure 6: *Multi-take video montages. (a) Two handheld takes of the same dialog and (b) two handheld takes of poetry recitation. (top) Cropped source (retimed) and target frames (left and right, resp.) with the region to be replaced marked in the first target frame. (bottom) Frames from the blended result that combine the target pose, background, and mouth with the source eyes and expression.*

the speech is the same in both source and target videos, it presents a problem for other situations when the motion of the target jaw and source mouth do not match. For these situations – changing the language during dubbing or in arbitrary face replacements – a full face model is necessary so that the source's jaw can also be transferred (Fig. 8 a).

Second, our experience using the original interior-only face model confirmed earlier psychological studies that had concluded that face shape is one of the stronger cues for identity. When source and target subjects differ, replacing the interior of the face was not always sufficient to convey the identity of the source subject, particularly when source and target face shapes differ significantly.

In Vlasic et al., face texture can come from either the source or the target, and morphable model parameters can be a mixture of source and target. When the target texture is used, as in their puppetry application, blending the warped texture is relatively easy. However, the expressiveness of the result stems exclusively from the morphable model, which is limited and lacks the detail and nuances of real facial performances in video. On the other hand, taking face texture from the source makes the task of blending far more difficult; as can be seen in Fig. 5, the naïve blending of source face texture into the target used in Vlasic et al. produces bleeding and flickering artifacts that are mitigated with our seam finding and blending method.

User study To quantitatively and objectively evaluate our system, we ran a user study using Amazon's Mechanical Turk. Our test set consisted of 24 videos: 10 unmodified videos, 10 videos with replaced faces, and four additional videos designed to verify that the subjects were watching the videos and not simply clicking on random responses. All videos were presented at 640×360 pixels for five seconds and then disappeared from the page to prevent the subject from analyzing the final frame.

The subjects were informed that the video they viewed was either "captured directly by a video camera" or "manipulated by a computer program." They were asked to respond to the statement "This video was captured directly by a video camera" by choosing a response from a five-point Likert scale: strongly agree (5), agree (4), neither agree nor disagree (3), disagree (2), or strongly disagree (1). We collected 40 distinct opinions per video and paid the subjects $\$0.04$ per opinion per video. The additional four videos began with similar footage as the rest but then instructed the subjects to click a specific response, e.g., 'agree', to verify that they were paying attention. Subjects who did not respond as instructed to these videos were discarded from the study. Approximately 20 opinions per video remained after removing these users.

The average response for the face-replaced videos was 4.1, indicating that the subjects believed the videos were captured directly by a camera and were not manipulated by a computer program. The average response for the authentic videos was 4.3, indicating a

Figure 7: *Dubbing. (top) Cropped source and target frames (left and right, resp.) from an indoor recording of dialog in English and an outdoor recording in Hindi, respectively. (bottom) Frames from the blended result. Differences in lighting and mouth/chin position between source and target are seamlessly combined in the result.*

slightly stronger belief that the videos were captured by a camera. None of the face-replaced videos had a median score below 4 and three of the videos had a median score of 5. These results indicate that our method can produce convincing videos that look similar to those coming directly from a camera.

Limitations Our approach is not without limitations (Fig. 8). Tracking is based on optical flow, which requires that the lighting change slowly over the face. High frequency lighting, such as hard shadows, must be avoided to ensure good tracking. Additionally, the method assumes an orthographic camera; while estimation of parameters of a more sophisticated camera model is possible, we use the simple model and shot our input videos with longer focal lengths that better approximate an orthographic projection. Finally, tracking often degrades beyond the range of poses outside $\pm 45^o$ from frontal. Even with successful tracking, the geometric fit can cause artifacts in the final result. For example, the fit is sometimes insufficient for the large pose differences between source and target. This is particularly noticeable in the nose area when, for example, the head is significantly tilted downwards, causing the nose to distort slightly.

Pose is also constrained to be sufficiently similar between source and target to prevent occluded regions in the source face from appearing in the pose-transformed target frame. For cases where we have control over source acquisition, the source subject can be captured in a frontal pose as we do here, or in a pose similar to the target, both ensuring no occluded regions. However when existing footage is used as the source, it is necessary to ensure compatible pose between source and target. This issue could be alleviated by automatic or user-assisted inpainting that derives the missing texture from spatially and temporally adjacent pixels in the video sequence.

In all examples shown here, source / target pairs are of the same gender and approximate age and thus of roughly similar proportions. Any difference in face shape can be accounted for by a single global scale to ensure the source face covers the target. For vastly different face shape, e.g., a child and adult, this may not be sufficient. However it is plausible to add a 2D warping step, similar to that used in [Jain et al. 2010], that warps the target face and nearby background to match the source before blending.

Lighting must also be similar between source and target. For multi-take montage scenarios, where source and target are typically captured in close succession in the same setting, this condition is trivially met. Likewise, when either the source or target is captured in a studio setting, with full control over the lighting setup, this condition can also be met with the same efforts required for plausible green screening. However such matching can be difficult for novices or may be impossible if the source and target are from existing footage.

Figure 8: *Failure cases. (a) Split frame of two nearby frames in a blended result where the model does not cover the full face. (b) When the tracking fails, the source content for replacement is distorted, seen here after alignment. (c) Significant differences in lighting between source and target lead to an unrealistic blended result, where the lighting on the right is darker on the source face but not in the target environment.*

Finally, seam finding and blending can fail for difficult inputs. For example, when hair falls along the forehead, there may be no seam that generates a natural blend between source and target. Strong differences in illuminations will lead to bleeding artifacts because it sometimes is not possible for the seam to avoid such regions. Fig. 8 shows some examples where these limitations are manifested in the final result.

8 Conclusions

We have presented a system for producing face replacements in video that requires only single-camera video and minimal user input and is robust under significant differences between source and target. We have shown with a user study that results generated with this method are perceived as realistic. Our method is useful in a va-

Figure 9: *Replacement. (top) Cropped source and target frames (left and right, resp.) showing casual conversation and head motion, with the target shot handheld. (bottom) Frames from the blended result, combining frames from two subjects with notably different expression, speech, pose, and face shape.*

riety of situations, including multi-take montage, dubbing, retargeting, and face replacement. Future improvements such as inpainting for occlusions during large pose variations, 2D background warping for vastly different face shapes, and lighting transfer between source and target will make this approach applicable to an even broader range of scenarios.

Acknowledgements

The authors thank Karen Xiao for help reprocessing the face scan data and Karen, Michelle Borkin, and Amanda Peters for participating as video subjects. This work was supported in part by the National Science Foundation under Grants No. PHY-0835713 and DMS-0739255.

References

AGARWALA, A., DONTCHEVA, M., AGRAWALA, M., DRUCKER, S., COLBURN, A., CURLESS, B., SALESIN, D., AND COHEN, M. 2004. Interactive digital photomontage. *ACM Trans. Graphics (Proc. SIGGRAPH) 23*, 3, 294–302.

ALEXANDER, O., ROGERS, M., LAMBETH, W., CHIANG, M., AND DEBEVEC, P. 2009. The digital emily project: Photoreal facial modeling and animation. In *ACM SIGGRAPH 2009 Courses*, 12:1–15.

BEELER, T., HAHN, F., BRADLEY, D., BICKEL, B., BEARDSLEY, P., GOTSMAN, C., SUMNER, B., AND GROSS, M. 2011 (to appear). High-quality passive facial performance capture using anchor frames. *ACM Trans. Graphics (Proc. SIGGRAPH) 3*, 27, 75:1–10.

BICKEL, B., BOTSCH, M., ANGST, R., MATUSIK, W., OTADUY, M., PFISTER, H., AND GROSS, M. 2007. Multi-scale capture of facial geometry and motion. *ACM Trans. Graphics (Proc. SIGGRAPH) 26*, 3, 33:1–10.

BITOUK, D., KUMAR, N., DHILLON, S., BELHUMEUR, P., AND NAYAR, S. K. 2008. Face swapping: Automatically replacing faces in photographs. *ACM Trans. Graphics (Proc. SIGGRAPH) 27*, 3, 39:1–8.

BLANZ, V., BASSO, C., POGGIO, T., AND VETTER, T. 2003. Reanimating faces in images and video. *Computer Graphics Forum 22*, 3, 641–650.

BLANZ, V., SCHERBAUM, K., VETTER, T., AND SEIDEL, H.-P. 2004. Exchanging faces in images. *Computer Graphics Forum (Proc. Eurographics) 23*, 3, 669–676.

BORSHUKOV, G., PIPONI, D., LARSEN, O., LEWIS, J., AND TEMPELAAR-LIETZ, C. 2003. Universal capture – Image-based facial animation for "The Matrix Reloaded". In *ACM SIGGRAPH 2003 Sketches & Applications*.

BOYKOV, Y., VEKSLER, O., AND ZABIH, R. 2001. Fast approximate energy minimization via graph cuts. *IEEE Trans. Pattern Analysis and Machine Intelligence 23*, 11, 1222–1239.

BRADLEY, D., HEIDRICH, W., POPA, T., AND SHEFFER, A. 2010. High resolution passive facial performance capture. *ACM Trans. Graphics (Proc. SIGGRAPH)*, 4, 41:1–10.

BREGLER, C., COVELL, M., AND SLANEY, M. 1997. Video Rewrite: Driving visual speech with audio. In *Proc. SIGGRAPH*, 353–360.

DECARLO, D., AND METAXAS, D. 1996. The integration of optical flow and deformable models with applications to human face shape and motion estimation. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 231–238.

ESSA, I., BASU, S., DARRELL, T., AND PENTLAND, A. 1996. Modeling, tracking and interactive animation of faces and heads: Using input from video. In *Proc. Computer Animation*, 68–79.

EVERINGHAM, M., SIVIC, J., AND ZISSERMAN, A. 2006. "Hello! My name is... Buffy" – automatic naming of characters in TV video. In *Proc. British Machine Vision Conference (BMVC)*, 899–908.

EZZAT, T., GEIGER, G., AND POGGIO, T. 2002. Trainable videorealistic speech animation. *ACM Trans. Graphics (Proc. SIGGRAPH) 21*, 3, 388–398.

FARBMAN, Z., HOFFER, G., LIPMAN, Y., COHEN-OR, D., AND LISCHINSKI, D. 2009. Coordinates for instant image cloning. *ACM Trans. Graphics (Proc. SIGGRAPH) 28*, 3, 67:1–9.

Figure 10: *Retargeting. (top) Cropped source (retimed) and target frames (left and right, resp.) from indoor recordings of poetry recitation. (bottom) Frames from the blended result combine the identity of the source with the background and timing of the target.*

FLAGG, M., NAKAZAWA, A., ZHANG, Q., KANG, S. B., RYU, Y. K., ESSA, I., AND REHG, J. M. 2009. Human video textures. In *Proc. Symp. Interactive 3D Graphics (I3D)*, 199–206.

GUENTER, B., GRIMM, C., WOOD, D., MALVAR, H., AND PIGHIN, F. 1998. Making faces. In *Proc. SIGGRAPH*, 55–66.

JAIN, A., THORMÄHLEN, T., SEIDEL, H.-P., AND THEOBALT, C. 2010. Moviereshape: Tracking and reshaping of humans in videos. *ACM Trans. Graphics (Proc. SIGGRAPH Asia) 29*, 5, 148:1–10.

JIA, J., SUN, J., TANG, C.-K., AND SHUM, H.-Y. 2006. Drag-and-drop pasting. *ACM Trans. Graphics (Proc. SIGGRAPH) 25*, 3, 631–637.

JONES, A., GARDNER, A., BOLAS, M., MCDOWALL, I., AND DEBEVEC, P. 2006. Simulating spatially varying lighting on a live performance. In *Proc. European Conf. Visual Media Production (CVMP)*, 127–133.

JOSHI, N., MATUSIK, W., ADELSON, E. H., AND KRIEGMAN, D. J. 2010. Personal photo enhancement using example images. *ACM Trans. Graphics 29*, 2, 12:1–15.

KEMELMACHER-SHLIZERMAN, I., SANKAR, A., SHECHTMAN, E., AND SEITZ, S. M. 2010. Being John Malkovich. In *Proc. European Conf. Computer Vision (ECCV)*, 341–353.

KWATRA, V., SCHÖDL, A., ESSA, I., TURK, G., AND BOBICK, A. 2003. Graphcut textures: Image and video synthesis using graph cuts. *ACM Trans. Graphics (Proc. SIGGRAPH) 22*, 3, 277–286.

LEYVAND, T., COHEN-OR, D., DROR, G., AND LISCHINSKI, D. 2008. Data-driven enhancement of facial attractiveness. *ACM Trans. Graphics (Proc. SIGGRAPH) 27*, 3, 38:1–9.

LI, H., ADAMS, B., GUIBAS, L. J., AND PAULY, M. 2009. Robust single-view geometry and motion reconstruction. *ACM Trans. Graphics (Proc. SIGGRAPH) 28*, 5, 175:1–10.

MA, W.-C., JONES, A., CHIANG, J.-Y., HAWKINS, T., FREDERIKSEN, S., PEERS, P., VUKOVIC, M., OUHYOUNG, M., AND DEBEVEC, P. 2008. Facial performance synthesis using deformation-driven polynomial displacement maps. *ACM Trans. Graphics (Proc. SIGGRAPH Asia) 27*, 5, 121:1–10.

PÉREZ, P., GANGNET, M., AND BLAKE, A. 2003. Poisson image editing. *ACM Trans. Graphics (Proc. SIGGRAPH) 22*, 3, 313–318.

PIGHIN, F. H., SZELISKI, R., AND SALESIN, D. 1999. Resynthesizing facial animation through 3d model-based tracking. In *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, 143–150.

RABINER, L., AND JUANG, B.-H. 1993. *Fundamentals of speech recognition*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

ROBERTSON, B. 2009. What's old is new again. *Computer Graphics World 32*, 1.

SINGULAR INVERSIONS INC., 2011. FaceGen Modeller manual. www.facegen.com.

SUNKAVALLI, K., JOHNSON, M. K., MATUSIK, W., AND PFISTER, H. 2010. Multi-scale image harmonization. *ACM Trans. Graphics (Proc. SIGGRAPH) 29*, 4, 125:1–10.

VIOLA, P. A., AND JONES, M. J. 2001. Robust real-time face detection. In *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, 747–755.

VLASIC, D., BRAND, M., PFISTER, H., AND POPOVIĆ, J. 2005. Face transfer with multilinear models. *ACM Trans. Graphics (Proc. SIGGRAPH) 24*, 3, 426–433.

WEISE, T., LI, H., GOOL, L. V., AND PAULY, M. 2009. Face/Off: Live facial puppetry. In *Proc. SIGGRAPH/Eurographics Symp. Computer Animation*, 7–16.

WILLIAMS, L. 1990. Performance-driven facial animation. *Computer Graphics (Proc. SIGGRAPH) 24*, 4, 235–242.

YANG, F., WANG, J., SHECHTMAN, E., BOURDEV, L., AND METAXAS, D. 2011. Expression flow for 3D-aware face component transfer. *ACM Trans. Graphics (Proc. SIGGRAPH) 27*, 3, 60:1–10.

ZHANG, L., SNAVELY, N., CURLESS, B., AND SEITZ, S. M. 2004. Spacetime faces: High resolution capture for modeling and animation. *ACM Trans. Graphics 23*, 3, 548–558.