Video Rewrite: Driving Visual Speech with Audio

Christoph Bregler, Michele Covell, Malcolm Slaney
Interval Research Corporation

ABSTRACT

Video Rewrite uses existing footage to create automatically new video of a person mouthing words that she did not speak in the original footage. This technique is useful in movie dubbing, for example, where the movie sequence can be modified to sync the actors' lip motions to the new soundtrack.

Video Rewrite automatically labels the phonemes in the training data and in the new audio track. Video Rewrite reorders the mouth images in the training footage to match the phoneme sequence of the new audio track. When particular phonemes are unavailable in the training footage, Video Rewrite selects the closest approximations. The resulting sequence of mouth images is stitched into the background footage. This stitching process automatically corrects for differences in head position and orientation between the mouth images and the background footage.

Video Rewrite uses computer-vision techniques to track points on the speaker's mouth in the training footage, and morphing techniques to combine these mouth gestures into the final video sequence. The new video combines the dynamics of the original actor's articulations with the mannerisms and setting dictated by the background footage. Video Rewrite is the first facial-animation system to automate all the labeling and assembly tasks required to resync existing footage to a new soundtrack.

CR Categories: I.3.3 [Computer Graphics]: Picture/Image Generation—Morphing; I.4.6 [Image Processing]: Segmentation—Feature Detection; I.3.8 [Computer Graphics]: Applications—Facial Synthesis; I.4.10 [Image Processing]: Applications—Feature Transformations.

Additional Keywords: Facial Animation, Lip Sync.

1 WHY AND HOW WE REWRITE VIDEO

We are very sensitive to the synchronization between speech and lip motions. For example, the special effects in *Forest Gump* are compelling because the Kennedy and Nixon footage is lip synched to the movie's new soundtrack. In contrast, close-ups in dubbed movies are often disturbing due to the lack of lip sync. Video Rewrite is a system for automatically synthesizing faces with proper lip sync. It can be used for dubbing movies, teleconferencing, and special effects.

1801 Page Mill Road, Building C, Palo Alto, CA, 94304. E-mail: bregler@cs.berkeley.edu, covell@interval.com, malcolm@interval.com. See the SIGGRAPH Video Proceedings or http://www.interval.com/papers/1997-012/ for the latest animations.

Video Rewrite automatically pieces together from old footage a new video that shows an actor mouthing a new utterance. The results are similar to labor-intensive special effects in *Forest Gump*. These effects are successful because they start from actual film footage and modify it to match the new speech. Modifying and reassembling such footage in a smart way and synchronizing it to the new sound track leads to final footage of realistic quality. Video Rewrite uses a similar approach but does not require labor-intensive interaction.

Our approach allows Video Rewrite to learn from example footage how a person's face changes during speech. We learn what a person's mouth looks like from a video of that person speaking normally. We capture the dynamics and idiosyncrasies of her articulation by creating a database of video clips. For example, if a woman speaks out of one side of her mouth, this detail is recreated accurately. In contrast, most current facial-animation systems rely on generic head models that do not capture the idiosyncrasies of an individual speaker.

To model a new person, Video Rewrite requires a small number (26 in this work) of hand-labeled images. This is the only human intervention that is required in the whole process. Even this level of human interaction is not a fundamental requirement: We could use face-independent models instead [Kirby90, Covell96].

Video Rewrite shares its philosophy with concatenative speech synthesis [Moulines90]. Instead of modeling the vocal tract, concatenative speech synthesis analyzes a corpus of speech, selects examples of phonemes, and normalizes those examples. Phonemes are the distinct sounds within a language, such as the /IY/ and /P/ in "teapot." Concatenative speech synthesizes new sounds by concatenating the proper sequence of phonemes. After the appropriate warping of pitch and duration, the resulting speech is natural sounding. This approach to synthesis is data driven: The algorithms analyze and resynthesize sounds using little hand-coded knowledge of speech. Yet they are effective at implicitly capturing the nuances of human speech.

Video Rewrite uses a similar approach to create new sequences of visemes. Visemes are the visual counterpart to phonemes. Visemes are visually distinct mouth, teeth, and tongue articulations for a language. For example, the phonemes /B/ and /P/ are visually indistinguishable and are grouped into a single viseme class.

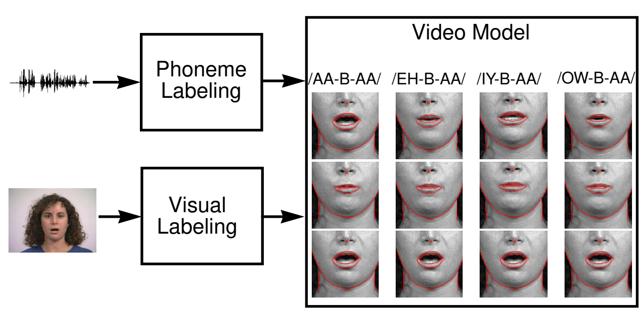

Figure 1: Overview of analysis stage. Video Rewrite uses the audio track to segment the video into triphones. Vision techniques find the orientation of the head, and the shape and position of the mouth and chin in each image. In the synthesis stage, Video Rewrite selects from this video model to synchronize new lip videos to any given audio.

Video Rewrite creates new videos using two steps: analysis of a training database and synthesis of new footage. In the *analysis* stage, Video Rewrite automatically segments into phonemes the audio track of the training database. We use these labels to segment the video track as well. We automatically track facial features in this segmented footage. The phoneme and facial labels together completely describe the visemes in the training database. In the *synthesis* stage, our system uses this video database, along with a new utterance. It automatically retrieves the appropriate viseme sequences, and blends them into a background scene using morphing techniques. The result is a new video with lip and jaw movements that synchronize to the new audio. The steps used in the analysis stage are shown in Figure 1; those of the synthesis stage are shown in Figure 2.

In the remainder of this paper, we first review other approaches to synthesizing talking faces (Section 2). We then describe the analysis and synthesis stages of Video Rewrite. In the analysis stage (Section 3), a collection of video is analyzed and stored in a database that matches sounds to video sequences. In the synthesis stage (Section 4), new speech is labeled, and the appropriate sequences are retrieved from the database. The final sections of this paper describe our results (Section 5), future work (Section 6), and contributions (Section 7).

2 SYNTHETIC VISUAL SPEECH

Facial-animation systems build a model of what a person's speech sounds and looks like. They use this model to generate a new output sequence, which matches the (new) target utterance. On the model-building side (analysis), there are typically three distinguishing choices: how the facial appearance is learned or described, how the facial appearance is controlled or labeled, and how the viseme labels are learned or described. For output-sequence generation (synthesis), the distinguishing choice is how the target utterance is characterized. This section reviews a representative sample of past research in these areas.

2.1 Source of Facial Appearance

Many facial-animation systems use a generic 3D mesh model of a face [Parke72, Lewis91, Guiard-Marigny94], sometimes adding texture mapping to improve realism [Morshima91, Cohen93, Waters95]. Another synthetic source of face data is hand-drawn images [Litwinowicz94]. Other systems use real faces for their source examples, including approaches that use 3D scans [Williams90] and still images [Scott94]. We use video footage to train Video Rewrite's models.

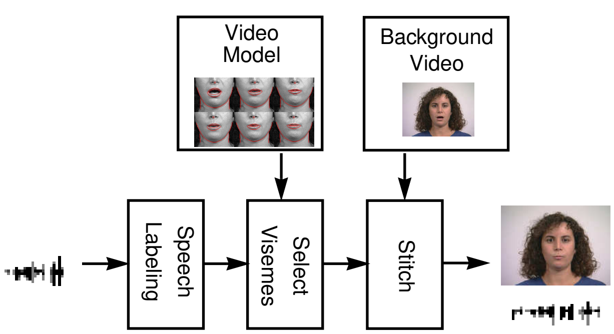

Figure 2: Overview of synthesis stage. Video Rewrite segments new audio and uses it to select triphones from the video model. Based on labels from the analysis stage, the new mouth images are morphed into a new background face.

2.2 Facial Appearance Control

Once a facial model is captured or created, the control parameters that exercise that model must be defined. In systems that rely on a 3D mesh model for appearance, the control parameters are the allowed 3D mesh deformations. Most of the image-based systems label the positions of specific facial locations as their control parameters. Of the systems that use facial-location labels, most rely on manual labeling of each example image [Scott94, Litwinowicz94]. Video Rewrite creates its video model by automatically labeling specific facial locations.

2.3 Viseme Labels

Many facial-animation systems label different visual configurations with an associated *phoneme*. These systems then match these phoneme labels with their corresponding labels in the target utterance. With synthetic images, the phoneme labels are artificial or are learned by analogy [Morshima91]. For natural images, taken from a video of someone speaking, the phonemic labels can be generated manually [Scott94] or automatically. Video Rewrite determines the phoneme labels automatically (Section 3.1).

2.4 Output-Sequence Generation

The goal of facial animation is to generate an image sequence that matches a target utterance. When phoneme labels are used, those for the target utterance can be entered manually [Scott94] or computed automatically [Lewis91, Morshima91]. Another option for phoneme labeling is to create the new utterance with synthetic speech [Parke72, Cohen93, Waters95]. Approaches, that do not use phoneme labels include motion capture of facial locations that are artificially highlighted [Williams90, Guiard-Marigny94] and manual control by an animator [Litwinowicz94]. Video Rewrite uses a combination of phoneme labels (from the target utterance) and facial-location labels (from the video-model segments). Video Rewrite derives all these labels automatically.

Video Rewrite is the first facial-animation system to automate all these steps and to generate realistic lip-synched video from natural speech and natural images.

3 ANALYSIS FOR VIDEO MODELING

As shown in Figure 1, the analysis stage creates an annotated database of example video clips, derived from unconstrained footage. We refer to this collection of annotated examples as a video model. This model captures how the subject's mouth and jaw move during speech. These training videos are labeled automatically with the phoneme sequence uttered during the video, and with the locations of fiduciary points that outline the lips, teeth, and jaw.

As we shall describe, the phonemic labels are from a time-aligned transcript of the speech, generated by a hidden Markov model (HMM). Video Rewrite uses the phonemic labels from the HMM to segment the input footage into short video clips, each showing three phonemes or a triphone. These triphone videos, with the fiduciary-point locations and the phoneme labels, are stored in the video model.

In Sections 3.1 and 3.2, we describe the visual and acoustic analyses of the video footage. In Section 4, we explain how to use this model to synthesize new video.

3.1 Annotation Using Image Analysis

Video Rewrite uses any footage of the subject speaking. As her face moves within the frame, we need to know the mouth position and the lip shapes at all times. In the synthesis stage, we use this information to warp overlapping videos such that they have the same lip shapes, and to align the lips with the background face.

Manual labeling of the fiduciary points around the mouth and jaw is error prone and tedious. Instead, we use computer-vision techniques to label the face and to identify the mouth and its shape. A major hurdle to automatic annotation is the low resolution of the images. In a typical scene, the lip region has a width of only 40 pixels. Conventional contour-tracking algorithms [Kass87, Yuille89] work well on high-contrast outer lip boundaries with some user interaction, but fail on inner lip boundaries at this resolution, due to the low signal-to-noise ratios. Grayscale-based algorithms, such as eigenimages [Kirby90, Turk91], work well at low resolutions, but estimate only the location of the lips or jaw, rather than estimating the desired fiduciary points. Eigenpoints [Covell96], and other extensions of eigenimages [Lanitis95], estimate control points reliably and automatically, even in such low-resolution images. As shown in Figure 3, eigenpoints learns how fiduciary points move as a function of the image appearance, and then uses this model to label new footage.

Video Rewrite labels each image in the training video using a total of 54 eigenpoints: 34 on the mouth (20 on the outer boundary, 12 on the inner boundary, 1 at the bottom of the upper teeth, and 1 at the top of the lower teeth) and 20 on the chin and jaw line. There are two separate eigenpoint analyses. The first eigenspace controls the placement of the 34 fiduciary points on the mouth, using 50×40 pixels around the nominal mouth location, a region that covers the mouth completely. The second eigenspace controls the placement of the 20 fiduciary points on the chin and jaw line, using 100×75 pixels around the nominal chin-location, a region that covers the upper neck and the lower part of the face.

We created the two eigenpoint models for locating the fiduciary points from a small number of images. We hand annotated only 26 images (of 14,218 images total; about 0.2%). We extended the hand-annotated dataset by morphing pairs of annotated images to form intermediate images, expanding the original 26 to 351 annotated images without any additional manual work. We then derived eigenpoints models using this extended data set.

We use eigenpoints to find the mouth and jaw and to label their contours. The derived eigenpoint models locate the facial features using six basis vectors for the mouth and six different vectors for the jaw. Eigenpoints then places the fiduciary points around the feature locations: 32 basis vectors place points around the lips and 64 basis vectors place points around the jaw.

Eigenpoints assumes that the features (the mouth or the jaw) are undergoing pure translational motion. It does a comparatively poor job at modeling rotations and scale changes. Yet, Video Rewrite is designed to use unconstrained footage. We expect rotations and scale changes. Subjects may lean toward the camera or turn away from it, tilt their heads to the side, or look up from under their eyelashes.

To allow for a variety of motions, we warp each face image into a standard reference plane, prior to eigenpoints labeling. We find the global transform that minimizes the mean-squared error between a large portion of the face image and a facial template. We currently use an affine transform [Black95]. The mask shown in Figure 4 defines the support of the minimization integral. Once the best global mapping is found, it is inverted and applied to the image, putting that face into the standard coordinate frame. We then perform eigenpoints analysis on this pre-warped image to find the fiduciary points. Finally, we back-project the fiduciary points through the global warp to place them on the original face image.

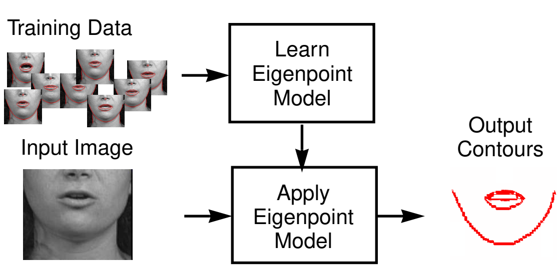

Figure 3: Overview of eigenpoints. A small set of hand-labeled facial images is used to train subspace models. Given a new image, the eigenpoint models tell us the positions of points on the lips and jaw.

The labels provided by eigenpoints allow us automatically to (1) build the database of example lip configurations, and (2) track the features in a background scene that we intend to modify. Section 4.2 describes how we match the points we find in step 1 to each other and to the points found in step 2.

3.2 Annotation Using Audio Analysis

All the speech data in Video Rewrite (and their associated video) are segmented into sequences of phonemes. Although single phonemes are a convenient representation for linguistic analysis, they are not appropriate for Video Rewrite. We want to capture the visual dynamics of speech. To do so correctly, we must consider *coarticulation*, which causes the lip shapes for many phonemes to be modified based on the phoneme's context. For example, the /T/ in "beet" looks different from the /T/ in "boot."

Therefore, Video Rewrite segments speech and video into triphones: collections of three sequential phonemes. The word "teapot" is split into the sequence of triphones /SIL-T-IY/ ,[1] /T-IY-P/ , /IY-P-AA/ , /P-AA-T/ , and /AA-T-SIL/ . When we synthesize a video, we emphasize the middle of each triphone. We cross-fade the overlapping regions of neighboring triphones. We thus ensure that the precise transition points are not critical, and that we can capture effectively many of the dynamics of both forward and backward coarticulation.

Video Rewrite uses HMMs [Rabiner89] to label the training footage with phonemes. We trained the HMMs using the TIMIT speech database [Lamel86], a collection of 4200 utterances with phonemic transcriptions that gives the uttered phonemes and their timing. Each of the 61 phoneme categories in TIMIT is modeled with a separate three-state HMM. The emission probabilities of each state are modeled with mixtures of eight Gaussians with diagonal covariances. For robustness, we split the available data by gender and train two speaker-independent, gender-specific systems, one based on 1300 female utterances, and one based on 2900 male utterances.

We used these gender-specific HMMs to create a fine-grained phonemic transcription of our input footage, using forced Viterbi

1. /SIL/ indicates silence. Two /SIL/ in a row are used at the beginnings and ends of utterances to allow all segments—including the beginning and end—to be treated as triphones.

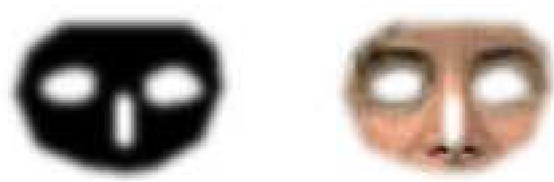

Figure 4: Mask used to estimate the global warp. Each image is warped to account for changes in the head's position, size, and rotation. The transform minimizes the difference between the transformed images and the face template. The mask (left) forces the minimization to consider only the upper face (right).

search [Viterbi67]. Forced Viterbi uses unaligned sentence-level transcriptions and a phoneme-level pronunciation dictionary to create a time-aligned phoneme-level transcript of the speech. From this transcript, Video Rewrite segments the video automatically into triphone videos, labels them, and includes them in the video model.

4 SYNTHESIS USING A VIDEO MODEL

As shown in Figure 2, Video Rewrite synthesizes the final lip-synced video by labeling the new speech track, selecting a sequence of triphone videos that most accurately matches the new speech utterance, and stitching these images into a background video.

The background video sets the scene and provides the desired head position and movement. The background sequence in Video Rewrite includes most of the subject's face as well as the scene behind the subject. The frames of the background video are taken from the source footage in the same order as they were shot. The head tilts and the eyes blink, based on the background frames.

In contrast, the different triphone videos are used in whatever order is needed. They simply show the motions associated with articulation. For all the animations in this paper, the triphone images include the mouth, chin, and part of the cheeks, so that the chin and jaw move and the cheeks dimple appropriately as the mouth articulates. We use illumination-matching techniques [Burt83] to avoid visible seams between the triphone and background images.

The first step in synthesis (Figure 2) is labeling the new soundtrack. We label the new utterance with the same HMM that we used to create the video-model phoneme labels. In Sections 4.1 and 4.2, we describe the remaining steps: selecting triphone videos and stitching them into the background.

4.1 Selection of Triphone Videos

The new speech utterance determines the target sequence of speech sounds, marked with phoneme labels. We would like to find a sequence of triphone videos from our database that matches this new speech utterance. For each triphone in the new utterance, our goal is to find a video example with exactly the transition we need, and with lip shapes that match the lip shapes in neighboring triphone videos. Since this goal often is not reachable, we compromise by a choosing a sequence of clips that approximates the desired transitions and shape continuity.

Given a triphone in the new speech utterance, we compute a matching distance to each triphone in the video database. The matching metric has two terms: the *phoneme-context distance*, D_p, and the *distance between lip shapes* in overlapping visual triphones, D_s. The total error is

$$\text{error} = \alpha D_p + (1-\alpha) D_s,$$

where the weight, α, is a constant that trades off the two factors.

The phoneme-context distance, D_p, is based on categorical distances between phoneme categories and between viseme classes. Since Video Rewrite does not need to create a new soundtrack (it needs only a new video track), we can cluster phonemes into viseme classes, based on their visual appearance.

We use 26 viseme classes. Ten are consonant classes: (1) /CH/, /JH/, /SH/, /ZH/; (2) /K/, /G/, /N/, /L/; (3) /T/, /D/, /S/, /Z/; (4) /P/, /B/, /M/; (5) /F/, /V/; (6) /TH/, /DH/; (7) /W/, /R/; (8) /HH/; (9) /Y/; and (10) /NG/. Fifteen are vowel classes: one each for /EH/, /EY/, /ER/, /UH/, /AA/ /AO/, /AW/, /AY/, /UW/, /OW/, /OY/, /IY/, /IH/, /AE/, /AH/. One class is for silence, /SIL/.

The phoneme-context distance, D_p, is the weighted sum of phoneme distances between the target phonemes and the video-model phonemes within the context of the triphone. If the phonemic categories are the same (for example, /P/ and /P/), then this distance is 0. If they are in different viseme classes (/P/ and /IY/), then the distance is 1. If they are in different phonemic categories but are in the same viseme class (/P/ and /B/), then the distance is a value between 0 and 1. The intraclass distances are derived from published confusion matrices [Owens85].

In D_p, the center phoneme of the triphone has the largest weight, and the weights drop smoothly from there. Although the video model stores only triphone images, we consider the triphone's original context when picking the best-fitting sequence. In current animations, this context covers the triphone itself, plus one phoneme on either side.

The second term, D_s, measures how closely the mouth contours match in overlapping segments of adjacent triphone videos. In synthesizing the mouth shapes for "teapot" we want the contours for the /IY/ and /P/ in the lip sequence used for /T-IY-P/ to match the contours for the /IY/ and /P/ in the sequence used for /IY-P-AA/. We measure this similarity by computing the Euclidean distance, frame by frame, between four-element feature vectors containing the overall lip width, overall lip height, inner lip height, and height of visible teeth.

The lip-shape distance (D_s) between two triphone videos is minimized with the correct time alignment. For example, consider the overlapping contours for the /P/ in /T-IY-P/ and /IY-P-AA/. The /P/ phoneme includes both a silence, when the lips are pressed together, and an audible release, when the lips move rapidly apart. The durations of the initial silence within the /P/ phoneme may be different. The phoneme labels do not provide us with this level of detailed timing. Yet, if the silence durations are different, the lip-shape distance for two otherwise-well-matched videos will be large. This problem is exacerbated by imprecision in the HMM phonemic labels.

We want to find the temporal overlap between neighboring triphones that maximizes the similarity between the two lip shapes. We shift the two triphones relative to each other to find the best temporal offset and duration. We then use this optimal overlap both in computing the lip-shape distance, D_s, and in cross-fading the triphone videos during the stitching step. The optimal overlap is the one that minimizes D_s while still maintaining a minimum-allowed overlap.

Since the fitness measure for each triphone segment depends on that segment's neighbors in both directions, we select the sequence of triphone segments using dynamic programming over the entire utterance. This procedure ensures the selection of the optimal segments.

4.2 Stitching It Together

Video Rewrite produces the final video by stitching together the appropriate entries from the video database. At this point, we have already selected a sequence of triphone videos that most closely matches the target audio. We need to align the overlapping lip images temporally. This internally time-aligned sequence of videos is then time aligned to the new speech utterance. Finally, the resulting sequences of lip images are spatially aligned and are stitched into the background face. We describe each step in turn.

4.2.1 Time Alignment of Triphone Videos

We have a sequence of triphone videos that we must combine to form a new mouth movie. In combining the videos, we want to maintain the dynamics of the phonemes and their transitions. We need to time align the triphone videos carefully before blending them. If we are not careful in this step, the mouth will appear to flutter open and closed inappropriately.

We align the triphone videos by choosing a portion of the overlapping triphones where the two lips shapes are as similar as possible. We make this choice when we evaluate D_s to choose the sequence of triphone videos (Section 4.1). We use the overlap duration and shift that provide the minimum value of D_s for the given videos.

4.2.2 Time Alignment of the Lips to the Utterance

We now have a self-consistent temporal alignment for the triphone videos. We have the correct articulatory motions, in the correct order to match the target utterance, but these articulations are not yet time aligned with the target utterance.

We align the lip motions with the target utterance by comparing the corresponding phoneme transcripts. The starting time of the center phone in the triphone sequence is aligned with the corresponding label in the target transcript. The triphone videos are then stretched or compressed such that they fit the time needed between the phoneme boundaries in the target utterance.

4.2.3 Combining of the Lips and the Background

The remaining task is to stitch the triphone videos into the background sequence. The correctness of the facial alignment is critical to the success of the recombination. The lips and head are constantly moving in the triphone and background footage. Yet, we need to align them all so that the new mouth is firmly planted on the face. Any error in spatial alignment causes the mouth to jitter relative to the face—an extremely disturbing effect.

We again use the mask from Figure 4 to help us find the optimal global transform to register the faces from the triphone videos with the background face. The combined tranforms from the mouth and background images to the template face (Section 3.1) give our starting estimate in this search. Re-estimating the global transform by directly matching the triphone images to the background improves the accuracy of the mapping.

We use a replacement mask to specify which portions of the final video come from the triphone images and which come from the background video. This replacement mask warps to fit the new mouth shape in the triphone image and to fit the jaw shape in the background image. Figure 5 shows an example replacement mask, applied to triphone and background images.

Local deformations are required to stitch the shape of the mouth and jaw line correctly. These two shapes are handled differently. The mouth's shape is completely determined by the triphone images. The only changes made to these mouth shapes are imposed to align the mouths within the overlapping triphone images: The lip shapes are linearly cross-faded between the shapes in the overlapping segments of the triphone videos.

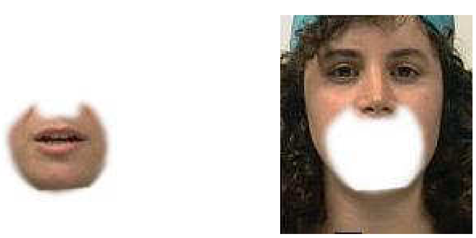

Figure 5: Facial fading mask. This mask determines which portions of the final movie frames come from the background frame, and which come from the triphone database. The mask should be large enough to include the mouth and chin. These images show the replacement mask applied to a triphone image, and its inverse applied to a background image. The mask warps according to the mouth and chin motions.

The jaw's shape, on the other hand, is a combination of the background jaw line and the two triphone jaw lines. Near the ears, we want to preserve the background video's jaw line. At the center of the jaw line (the chin), the shape and position are determined completely by what the mouth is doing. The final image of the jaw must join smoothly together the motion of the chin with the motion near the ears. To do this, we smoothly vary the weighting of the background and triphone shapes as we move along the jawline from the chin towards the ears.

The final stitching process is a three-way tradeoff in shape and texture among the fade-out lip image, the fade-in lip image, and the background image. As we move from phoneme to phoneme, the relative weights of the mouth shapes associated with the overlapping triphone-video images are changed. Within each frame, the relative weighting of the jaw shapes contributed by the background image and of the triphone-video images are varied spatially.

The derived fiduciary positions are used as control points in morphing. All morphs are done with the Beier-Neely algorithm [Beier92]. For each frame of the output image we need to warp four images: the two triphones, the replacement mask, and the background face. The warping is straightforward since we automatically generate high-quality control points using the eigenpoints algorithm.

5 RESULTS

We have applied Video Rewrite to several different training databases. We recorded one video dataset specifically for our evaluations. Section 5.1 describes our methods to collect this data and create lip-sync videos. Section 5.2 evaluates the resulting videos.

We also trained video models using truncated versions of our evaluation database. Finally, we used old footage of John F. Kennedy. We present the results from these experiments in Section 5.3.

5.1 Methods

We recorded about 8 minutes of video, containing 109 sentences, of a subject narrating a fairy tale. During the reading, the subject was asked to directly face the camera for some parts (still-head video) and to move and glance around naturally for others (moving-head video). We use these different segments to study the errors in local deformations separately from the errors in global spatial registration. The subject was also asked to wear a hat during the filming. We use this landmark to provide a quantitative evaluation of our global alignment. The hat is strictly outside all our alignment masks and our eigenpoints models. Thus, having the subject wear the hat does not effect the magnitude or type of errors that we expect to see in the animations—it simply provides us with a reference marker for the position and movement of her head.

To create a video model, we trained the system on all the still-head footage. Video Rewrite constructed and annotated the video model with just under 3500 triphone videos automatically, using HMM labeling of triphones and eigenpoint labeling of facial contours.

Video Rewrite was then given the target sentence, and was asked to construct the corresponding image sequence. To avoid unduly optimistic results, we removed from the database the triphone videos from training sentences similar to the target. A training sentence was considered similar to the target if the two shared a phrase two or more words long. Note that Video Rewrite would not normally pare the database in this manner: Instead, it would take advantage of these coincidences. We remove the similar sentences to avoid biasing our results.

We evaluated our output footage both qualitatively and quantitatively. Our qualitative evaluation was done informally, by a panel of observers. There are no accepted metrics for evaluating lip-synced footage. Instead, we were forced to rely on the qualitative judgements listed in Section 5.2.

Only the (global) spatial registration is evaluated quantitatively. Since our subject wore a hat that moved rigidly with her upper head, we were able to measure quantitatively our global-registration error on this footage. We did so by first warping the full frame (instead of just the mouth region) of the triphone image into the coordinate frame of the background image. If this global transformation is correct, it should overlay the two images of the hat exactly on top of one another. We measured the error by finding the offset of the correlation peak for the image regions corresponding to the front of the hat. The offset of the peak is the registration error (in pixels).

5.2 Evaluation

Examples of our output footage can be seen at http://www.interval.com/papers/1997-012/. The top row of Figure 6 shows example frames, extracted from these videos. This section describes our evaluation criteria and the results.

5.2.1 Lip and Utterance Synchronization

How well are the lip motions synchronized with the audio? We evaluate this measure on the still-head videos. There occasionally are visible timing errors in plosives and stops.

5.2.2 Triphone-Video Synchronization

Do the lips flutter open and closed inappropriately? This artifact usually is due to synchronization error in overlapping triphone videos. We evaluated this measure on the still-head videos. We do not see any artifacts of this type.

5.2.3 Natural Articulation

Assuming that neither of the artifacts from Sections 5.2.1 or 5.2.2 appear, do the lip and teeth articulations look natural? Unnatural-looking articulation can result if the desired sequence of phonemes is not available in the database, and thus another sequence is used in its place. In our experiments, this replacement occurred on 31 percent of the triphone videos. We evaluated this measure on the still-head videos. We do not see this type of error when we use the full video model. Additional experiments in this area are described in Section 5.3.1.

5.2.4 Fading-Mask Visibility and Extent

Does the fading mask show? Does the animation have believable texture and motion around the lips and chin? Do the dimples move in sync with the mouth? We evaluated this measure on all the output videos. The still-head videos better show errors associated with the extent of the fading mask, whereas the moving-head videos better show errors due to interactions between the fading mask and the global transformation. Without illumination correction, we see artifacts in some of the moving-head videos, when the subject looked down so that the lighting on her face changed significantly. These artifacts disappear with adaptive illumination correction [Burt83].

5.2.5 Background Warping

Do the outer edges of the jaw line and neck, and the upper portions of the cheeks look realistic? Artifacts in these areas are due to incorrect warping of the background image or to a mismatch between the texture and the warped shape of the background image. We evaluated this measure on all the output videos. In some segments, we found minor artifacts near the outer edges of the jaw.

5.2.6 Spatial Registration

Does the mouth seem to float around on the face? Are the teeth rigidly attached to the skull? We evaluated this measure on the moving-head videos. No registration errors are visible.

We evaluated this error quantitatively as well, using the hat-registration metric described in Section 5.1. The mean, median, and maximum errors in the still-head videos were 0.6, 0.5, and 1.2 pixels (standard deviation 0.3); those in the moving-head videos were 1.0, 1.0, and 2.0 pixels (standard deviation 0.4). For comparison, the face covers approximately 85×120 pixels.

5.2.7 Overall Quality

Is the lip-sync believable? We evaluated this measure on all the output videos. We judged the overall quality as excellent.

Figure 6: Examples of synthesized output frames. These frames show the quality of our output after triphone segments have been stitched into different background video frames.

5.3 Other Experiments

In this section, we examine our performance using steadily smaller training databases (Section 5.3.1) and using historic footage (Section 5.3.2).

5.3.1 Reduction of Video Model Size

We wanted to see how the quality fell off as the number of data available in the video model were reduced. With the 8 minutes of video, we have examples of approximately 1700 different triphones (of around 19,000 naturally occurring triphones); our animations used triphones other than the target triphones 31 percent of the time. What happens when we have only 1 or 2 minutes of data? We truncated our video database to one-half, one-quarter, and one-eighth of its original size, and then reanimated our target sentences. The percent of mismatched triphones increased by about 15 percent with each halving of the database (that is, 46, 58, and 74 percent of the triphones were replaced in the reduced datasets). The perceptual quality also degraded smoothly as the database size was reduced. The video from the reduced datasets are shown on our web site.

5.3.2 Reanimation of Historic Footage

We also applied Video Rewrite to public-domain footage of John F. Kennedy. For this application, we digitized 2 minutes (1157 triphones) of Kennedy speaking during the Cuban missile crisis. Forty-five seconds of this footage are from a close-up camera, about 30 degrees to Kennedy's left. The remaining images are medium shots from the same side. The size ratio is approximately 5:3 between the close-up and medium shots. During the footage, Kennedy moves his head about 30 degrees vertically, reading his speech from notes on the desk and making eye contact with a center camera (which we do not have).

We used this video model to synthesize new animations of Kennedy saying, for example, “Read my lips” and “I never met Forrest Gump.” These animations combine the footage from both camera shots and from all head poses. The resulting videos are shown on our web site. The bottom row of Figure 6 shows example frames, extracted from these videos.

In our preliminary experiments, we were able to find the correct triphone sequences just 6% of the time. The lips are reliably synchronized to the utterance. The fading mask is not visible, nor is the background warping. However, the overall animation quality is not as good as our earlier results. The animations include some lip fluttering, because of the mismatched triphone sequences.

Our quality is limited for two reasons. The available viseme footage is distributed over a wide range of vertical head rotations. If we choose triphones that match the desired pose, then we cannot find good matches for the desired phoneme sequence. If we choose triphones that are well matched to the desired phoneme sequence, then we need to dramatically change the pose of the lip images. A large change in pose is difficult to model with our global (affine) transform. The lip shapes are distorted because we assumed, implicitly in the global transform, that the lips lie on a flat plane. Both the limited-triphone and pose problems can be avoided with additional data.

6 FUTURE WORK

There are many ways in which Video Rewrite could be extended and improved. The phonemic labeling of the triphone and background footage could consider the mouth- and jaw-shape information, as well as acoustic data [Bregler95]. Additional lip-image data and multiple eigenpoints models could be added, allowing larger out-of-plane head rotations. The acoustic data could be used in selecting the triphone videos, because facial expressions affect voice qualities (you can hear a smile). The synthesis could be made real-time, with low-latency.

In Sections 6.1 through 6.3, we explore extensions that we think are most promising and interesting.

6.1 Alignment Between Lips and Target

We currently use the simplest approach to time aligning the lip sequences with the target utterance: We rely on the phoneme boundaries. This approach provides a rough alignment between the motions in the lip sequence and the sounds in the target utterance. As we mentioned in Section 4.1, however, the phoneme boundaries are both imprecise (the HMM alignment is not perfect) and coarse (significant visual and auditory landmarks occur within single phonemes).

A more accurate way to time align the lip motions with the target utterance uses dynamic time warping of the audio associated with each triphone video to the corresponding segment of the target utterance. This technique would allow us to time align the auditory landmarks from the triphone videos with those of the target utterance, even if the landmarks occur at subphoneme resolution. This time alignment, when applied to the triphone image sequence, would then align the visual landmarks of the lip sequence with the auditory landmarks of the target utterance.

The overlapping triphone videos would provide overlapping and conflicting time warpings. Yet we want to keep fixed the time alignment of the overlapping triphone videos, as dictated by the visual distances (Section 4.1 and 4.2). Research is needed in how best to trade off these potentially conflicting time-alignment maps.

6.2 Animation of Facial Features

Another promising extension is animation of other facial parts, based on simple acoustic features or other criteria. The simplest version of this extension would change the position of the eyebrows with pitch [Ohala94]. A second extension would index the video model by both triphone and expression labels. Using such labels, we would select smiling or frowning lips, as desired. Alternatively, we could impose the desired expression on a neutral mouth shape, for those times when the appropriate combinations of triphones and expression are not available. To do this imposition correctly, we must separate which deformations are associated with articulations, and which are associated with expressions, and how the two interact. This type of factorization must be learned from examples [Tenenbaum97].

6.3 Perception of Lip Shapes

In doing this work, we solved many problems—automatic labeling, matching, and stitching—yet we found many situations where we did not have sufficient knowledge of how people perceive speaking faces. We would like to know more about how important the correct lip shapes and motions are in lip synching. For example, one study [Owens85] describes the confusibility of consonants in vowel–consonant–vowel clusters. The clustering of consonants into viseme class depends on the surrounding vowel context. Clearly, we need more sophisticated distance metrics within and between viseme classes.

7 CONTRIBUTIONS

Video Rewrite is a facial animation system that is driven by audio input. The output sequence is created from real video footage. It combines background video footage, including natural facial movements (such as eye blinks and head motions) with natural footage of mouth and chin motions. Video Rewrite is the first facial-animation system to automate all the audio- and video-labeling tasks required for this type of reanimation.

Video Rewrite can use images from unconstrained footage both to create the video model of the mouth and chin motions and to provide a background sequence for the final output footage. It preserves the individual characteristics of the subject in the original footage, even while the subject appears to mouth a completely new utterance. For example, the temporal dynamics of John F. Kennedy's articulatory motions can be preserved, reorganized, and reimposed on Kennedy's face.

Since Video Rewrite retains most of the background frame, modifying only the mouth area, it is well suited to applications such as movie dubbing. The setting and action are provided by the background video. Video Rewrite maintains an actor's visual mannerisms, using the dynamics of the actor's lips and chin from the video model for articulatory mannerisms, and using the background video for all other mannerisms. It maintains the correct timing, using the action as paced by the background video and speech as paced by the new soundtrack. It undertakes the entire process without manual intervention. The actor convincingly mouths something completely new.

ACKNOWLEDGMENTS

Many colleagues helped us. Ellen Tauber and Marc Davis graciously submitted to our experimental manipulation. Trevor Darrell and Subutai Ahmad contributed many good ideas to the algorithm development. Trevor, Subutai, John Lewis, Bud Lassiter, Gaile Gordon, Kris Rahardja, Michael Bajura, Frank Crow, Bill Verplank, and John Woodfill helped us to evaluate our results and the description. Bud Lassiter and Chris Seguine helped us with the video production. We offer many thanks to all.

REFERENCES

[Beier92] T. Beier, S. Neely. Feature-based image metamorphosis. *Computer Graphics,* 26(2):35–42, 1992. ISSN 0097-8930.

[Black95] M.J. Black, Y. Yacoob. Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motion. *Proc. IEEE Int. Conf. Computer Vision,* Cambridge, MA, pp. 374–381, 1995. ISBN 0-8186-7042-8.

[Bregler95] C. Bregler, S. Omohundro. Nonlinear manifold learning for visual speech recognition. *Proc. IEEE Int. Conf. Computer Vision,* Cambridge, MA, pp. 494–499, 1995. ISBN 0-8186-7042-8.

[Burt83] P.J. Burt, E.H. Adelson. A multiresolution spline with application to image mosaics. *ACM Trans. Graphics,* 2(4): 217–236, 1983. ISSN 0730-0301.

[Cohen93] M.M. Cohen, D.W. Massaro. Modeling coarticulation in synthetic visual speech. In *Models and Techniques in Computer Animation,* ed. N.M Thalman, D. Thalman, pp. 139–156, Tokyo: Springer-Verlag, 1993. ISBN 0-3877-0124-9.

[Covell96] M. Covell, C. Bregler. Eigenpoints. *Proc. Int. Conf. Image Processing,* Lausanne, Switzerland, Vol. 3, pp. 471–474, 1996. ISBN 0-7803-3258-x.

[Guiard-Marigny94] T. Guiard-Marigny, A. Adjoudani, C. Benoit. A 3-D model of the lips for visual speech synthesis. *Proc. ESCA/IEEE Workshop on Speech Synthesis*, New Paltz, NY, pp. 49–52, 1994.

[Kass87] M. Kass, A. Witkin, D. Terzopoulos. Snakes: Active contour models. *Int. J. Computer Vision,* 1(4):321–331, 1987. ISSN 0920-5691.

[Kirby90] M. Kirby, L. Sirovich. Application of the Karhunen-Loeve procedure for the characterization of human faces. *IEEE PAMI*, 12(1):103–108, Jan. 1990. ISSN 0162-8828.

[Lamel86] L. F. Lamel, R. H. Kessel, S. Seneff. Speech database development: Design and analysis of the acoustic-phonetic corpus. *Proc. Speech Recognition Workshop (DARPA)*, Report #SAIC-86/1546, pp. 100–109, McLean VA: Science Applications International Corp., 1986.

[Lanitis95] A. Lanitis, C.J. Taylor, T.F. Cootes. A unified approach for coding and interpreting face images. *Proc. Int. Conf. Computer Vision,* Cambridge, MA, pp. 368–373, 1995. ISBN 0-8186-7042-8.

[Lewis91] J.Lewis. Automated lip-sync: Background and techniques. *J.Visualization and Computer Animation,* 2(4):118–122, 1991. ISSN 1049-8907.

[Litwinowicz94] P. Litwinowicz, L. Williams. Animating images with drawings. *SIGGRAPH 94*, Orlando, FL, pp. 409–412, 1994. ISBN 0-89791-667-0.

[Morishima91] S. Morishima, H. Harashima. A media conversion from speech to facial image for intelligent man-machine interface. *IEEE J Selected Areas Communications,* 9 (4):594–600, 1991. ISSN 0733-8716.

[Moulines90] E. Moulines, P. Emerard, D. Larreur, J. L. Le Saint Milon, L. Le Faucheur, F. Marty, F. Charpentier, C. Sorin. A real-time French text-to-speech system generating high-quality synthetic speech. *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, Albuquerque, NM, pp. 309–312, 1990.

[Ohala94] J.J. Ohala. The frequency code underlies the sound symbolic use of voice pitch. In *Sound Symbolism*, ed. L. Hinton, J. Nichols, J. J. Ohala, pp. 325–347, Cambridge UK: Cambridge Univ. Press, 1994. ISBN 0-5214-5219-8.

[Owens85] E. Owens, B. Blazek. Visemes observed by hearing-impaired and normal-hearing adult viewers. *J. Speech and Hearing Research*, 28:381–393, 1985. ISSN 0022-4685.

[Parke72] F. Parke. Computer generated animation of faces. *Proc. ACM National Conf.*, pp. 451–457, 1972.

[Rabiner89] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. In *Readings in Speech Recognition*, ed. A. Waibel, K. F. Lee, pp. 267–296, San Mateo, CA: Morgan Kaufmann Publishers, 1989. ISBN 1-5586-0124-4.

[Scott94] K.C. Scott, D.S. Kagels, S.H. Watson, H. Rom, J.R. Wright, M. Lee, K.J. Hussey. Synthesis of speaker facial movement to match selected speech sequences. *Proc. Australian Conf. Speech Science and Technology*, Perth Australia, pp. 620–625, 1994. ISBN 0-8642-2372-2.

[Tenenbaum97] J. Tenenbaum, W. Freeman. Separable mixture models: Separating style and content. In *Advances in Neural Information Processing 9,* ed. M. Jordan, M. Mozer, T. Petsche, Cambridge, MA: MIT Press, (in press).

[Turk91] M. Turk, A. Pentland. Eigenfaces for recognition. *J. Cognitive Neuroscience,* 3(1):71–86, 1991. ISSN 0898-929X

[Viterbi67] A. J. Viterbi. Error bounds for convolutional codes and an asymptotically optimal decoding algorithm. *IEEE Trans. Informat. Theory,* IT-13:260–269, 1967. ISSN 0018-9448.

[Waters95] K. Waters, T. Levergood. DECface: A System for Synthetic Face Applications. *J. Multimedia Tools and Applications,* 1 (4):349–366, 1995. ISSN 1380-7501.

[Williams90] L. Williams. Performance-Driven Facial Animation. *Computer Graphics* (Proceedings of SIGGRAPH 90), 24(4):235–242, 1990. ISSN 0097-8930.

[Yuille89] A.L. Yuille, D.S. Cohen, P.W. Hallinan. Feature extraction from faces using deformable templates. *Proc. IEEE Computer Vision and Pattern Recognition,* San Diego, CA, pp. 104–109, 1989. ISBN 0-8186-1952-x.