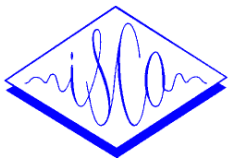

ESCA Workshop on
Audio-Visual Speech Processing
Rhodes, Greece
September 26-27, 2997

Video Rewrite: Visual Speech Synthesis from Video[1]

Christoph Bregler, Michele Covell, Malcolm Slaney
Interval Research Corporation
1801 Page Mill Road, Building C
Palo Alto, CA 94304 USA
http://www.interval.com/papers/1997-022

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.

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.

1 INTRODUCTION

Humans are extremely sensitive to the synchronization between speech and lip motions. For example, the special effects in *Forrest 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. Our system, Video Rewrite, automatically synthesizes faces with proper lip sync. It can be used for dubbing movies, teleconferencing, and special effects.

Video Rewrite automatically pieces together from old footage a new video that shows an actor mouthing a new utterance. The results are similar to the labor-intensive special effects in *Forrest Gump* and to the Actors system from JPL [Scott94]. Video Rewrite learns 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. In contrast, most current facial-animation systems rely on generic head models that do not capture the idiosyncrasies of an individual speaker (see, for example, [Parke72]).

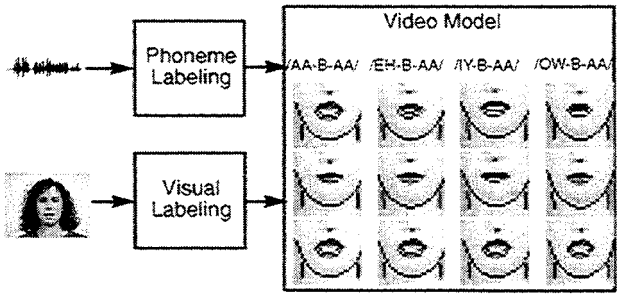

Figure 1: Overview of analysis stage. Video Rewrite uses the audio track to segment the video into triphones. Vision techniques find the orientation of the head, and the shape and position of the mouth and chin, in each image. This video model is used in the synthesis stage.

1. This paper is based on a previous paper [Bregler97]. Our previous paper describes the related literature and presents more details about our evaluation techniques, contributions, and future plans.

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. Concatenative speech synthesis creates 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 creates new videos using two steps: analysis of a training database and synthesis of new footage. In the *analysis* stage, Video Rewrite automatically segments the training database into phonemes. The phonemes and automatically tracked facial labels completely describe the visemes in the training database. In the *synthesis* stage, our system selects from this video database, as dictated by 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

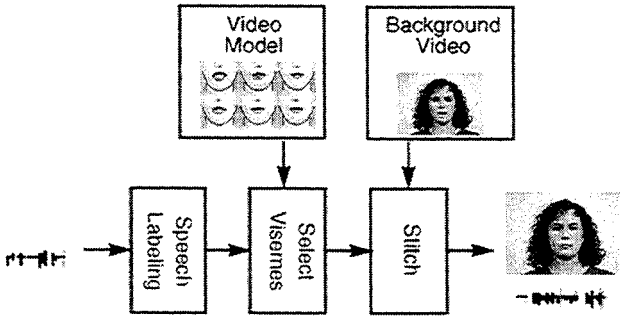

Figure 2: Overview of synthesis stage. Video Rewrite segments new audio and uses these segments to select triphones from the video model. Based on labels from the analysis stage, the new mouth images are morphed into a new background face.

are shown in Figure 1, and are described in Section 2. Those of the synthesis stage are shown in Figure 2, and are described in Section 3. Section 4 summarizes our results.

2 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.

In Sections 2.1 and 2.2, we describe the visual and acoustic analyses of the video footage. In Section 3, we explain how we use this model to synthesize new video.

2.1 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.

The eigenpoints algorithm [Covell96] locates fiduciary points in images; like related techniques [Lanitis95, Beymer93], it works reliably and automatically, even in low-resolution images. We use the eigenpoints algorithm to delineate the mouth and jaw.

We created two eigenpoints models for locating the fiduciary points from a small number of images. We hand annotated only 20 images (of 3654 images total; about 0.5 percent). We extended the hand-annotated dataset by morphing pairs of annotated images to form intermediate images, expanding the original 20 to 210 annotated images without any additional manual work. We then derived the two eigenpoints models using this extended data set.

The derived eigenpoints models locate the facial features using six basis vectors for the mouth and six different vectors for the jaw. The eigenpoints models place the fiduciary points around the feature locations: 32 basis vectors place points around the lips, and 64 basis vectors place points around the jaw.

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 complete this minimization using the areas around the forehead, eyebrows, upper cheeks and lower nose. We currently use an ellipsoidal transform [Basu96], followed by an affine transform [Black95]. The ellipsoid allows us to describe the curvature of the face and to compensate for changes in pose. Subsequent processing using an affine transform provides more accurate and reliable estimates of the head's translation and rotation. Once the global mapping with the minimum mean square error 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 prewarped image to find the fiduciary points.

The labels provided by eigenpoints allow us automatically (1) to build the database of example lip configurations, and (2) to track the features in a background scene that we intend to modify. Section 3.2 describes how we match the points that we find in step 1 to one another and to the points found in step 2.

2.2 Annotation Using Audio Analysis

All the speech data in Video Rewrite are segmented into triphones. Video Rewrite uses these labels to segment the video. When we synthesize a new video, 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 coarticulation.

We used gender-specific HMMs, trained on TIMIT data, to create a fine-grained phonemic transcription of our input footage, using forced Viterbi search. From this transcript, Video Rewrite segments the video automatically into triphone videos, labels them, and includes them in the video model.

3 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 the sequence of triphone videos that most accurately matches the new speech utterance, and stitching these images into a background video.

The background video 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. We use illumination-matching techniques [Burt83] to avoid visible seams between the triphone and background images.

Labeling the new soundtrack is the first step in synthesis (Figure 2). 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.

3.1 Selecting Triphone Videos

The new speech utterance, marked with phoneme labels, determines the target sequence of lip shapes. 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 that we need. 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 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 of the target and the video-model phonemes are the same (for example, /P/ and /P/), then this distance is 0. If the target and the video-model phonemes 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 the phoneme-context distance, D_p, the center phoneme of the triphone has the largest weight, and the weights drop smoothly on either side. Although the video model stores only triphone images, we consider the triphone's larger 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 shapes match in overlapping segments of adjacent triphone videos. In synthesizing the mouth shapes for "teapot," we want the shapes for the /IY/ and /P/ in the lip sequence used for /T-IY-P/ to match the shapes 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 shapes for the /P/ in /T-IY-P/ and /IY-P-AA/. 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.

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.

3.2 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 the 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 aligned spatially and are stitched into the background face. We describe each step in turn.

3.2.1 Time aligning the triphones

We have a sequence of triphone videos that we must combine to form a new mouth movie. 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 using the overlap duration and shift that provide the minimum value of D_s for the given videos.

We then 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 so that they fit the time needed between the phoneme boundaries in the target utterance.

3.2.2 Combining the lips and the background

The remaining task is to stitch the triphone videos into the background sequence. 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 use the combined tranforms from the mouth and background images to the template face (Section 2.1) as our starting estimate for this alignment. We improve its accuracy by reestimating the global transform, directly matching the triphone images to the background face.

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. The mask replaces the mouth, chin, and smile lines.

The mouth's shape is completely determined by the triphone images. When the triphone sequences overlap in time, their mouth shapes are aligned with one another: The mouth shapes are linearly cross-faded between the shapes in the overlapping segments of the triphone videos.

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. In between, we smoothly vary the weighting of the background and triphone shapes along the jawline.

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 generate high-quality control points automatically using the eigenpoints algorithm.

4 RESULTS

We applied Video Rewrite to public-domain footage of former President 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, positioned about 30 degrees to Kennedy's left. The remaining images are medium-distance 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 20 degrees vertically, reading his speech from notes on the desk and making eye contact with a center camera (footage from 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 at our web site, http://www.interval.com/papers/1997-022/. Figure 3 shows example frames, extracted from these videos.

We evaluated our Kennedy results qualitatively along the following dimensions: synchronization between lip videos and between the composite lips and the utterance; spatial registration between the lip videos and between the composite lips and the background head; quality of the illumination matching between the lips and the background head; visibility of the chosen fading-mask extent and of the background warping; naturalness of the composited articulation; and the overall quality of the video.

- There are visible timing errors in 1 percent of the phonemes. These timing errors all occur during plosives and stops. There are no visible artifacts due to synchronization errors between triphone videos.
- The lips are distorted unnaturally in 8 percent of the output frames. This distortion is caused by mistakes in the estimate of out-of-plane facial curvature. We see no other errors in the alignment between the lips and the background face.
- The illumination matching is accurate. There are no visible artifacts from illumination mismatches.
- The fading mask occasionally includes nonfacial regions (e.g., the flag behind Kennedy or the President's shirt collar). This error results in visible artifacts in 4 percent of the output frames, when lips from one head pose are warped into another pose.
- Unnatural-looking articulation results occasionally from replacement of a desired (but unavailable) triphone sequence. In our experiments with Kennedy, this type of replacement occurs on 94 percent of the triphone videos. Of those replacements, 4 percent are judged unnatural looking.
- Despite the foregoing occasional artifacts, the overall quality of the final video is judged as excellent.

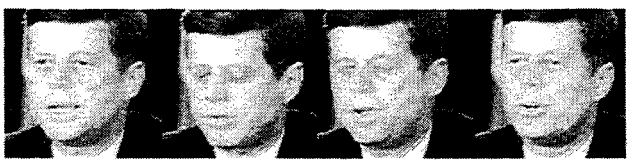

Figure 3: Examples of synthesized output frames. These frames show the quality of our output after triphone segments have been stitched into different background video frames.

5 CONCLUSION

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.

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, Lyn Dupré, 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

Basu, S., Essa, I., & Pentland, A. (1996), "Motion regularization for model-based head tracking", *Proc. Int. Conf. Pattern Recognition*, Vienna, Austria, pp. 611–616.

Beier, T., & Neely, S. (1992), "Feature-based image metamorphosis", *Computer Graphics*, 26(2): 35–42.

Beymer, D., Shahsua, A., & Poggio, T. (1993), "Example-based image analysis and synthesis", Memo No. 1431, AI Lab, MIT.

Black, M. J., & Yacoob, Y. (1995), "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.

Bregler, C., Covell, M., & Slaney, M. (1997), "Video Rewrite: Driving visual speech with audio", *SIGGRAPH 97*, Los Angeles, CA (in press).

Burt, P. J., & Adelson, E. H. (1983), "A multiresolution spline with application to image mosaics", *ACM Trans. Graphics*, 2(4): 217–236.

Covell, M., & Bregler, C. (1996), "Eigenpoints", *Proc. Int. Conf. Image Processing*, Lausanne, Switzerland, Vol. 3, pp. 471–474.

Lanitis, A., Taylor, C.J., & Cootes, T.F. (1995), "A unified approach for coding and interpreting face images", *Proc. Int. Conf. Computer Vision*, Cambridge, MA, pp. 368–373.

Moulines, E., et al. (1990), "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.

Owens, E., & Blazek, B. (1985), "Visemes observed by hearing-impaired and normal-hearing adult viewers", *J. Speech and Hearing Research*, 28(3): 381–393.

Parke, F. (1972), "Computer generated animation of faces", *Proc. ACM National Conf.*, Boston, MA, pp. 451–457.

Scott,K.C., et al. (1994), "Synthesis of speaker facial movement to match selected speech sequences", *Proc. Australian Conf. Speech Science and Technology*, Perth, Australia, pp. 620–625.