VIDEOREALISTIC TALKING FACES: A MORPHING APPROACH

Tony Ezzat *Tomaso Poggio*
tonebone@ai.mit.edu tp-temp@ai.mit.edu
MIT Center for Biological and Computational Learning
45 Carleton Street E25-201
Cambridge, MA 02141

ABSTRACT

We present a method for the construction of a videorealistic text-to-audiovisual speech synthesizer. A visual corpus of a subject enunciating a set of key words is initially recorded. The key words are chosen so that they collectively contain most of the American English viseme images, which are subsequently identified and extracted from the data by hand. Next, using optical flow methods borrowed from the computer vision literature, we compute realistic transitions between every viseme to every other viseme. The images along these transition paths are generated using a morphing method. Finally, we exploit phoneme and timing information extracted from a text-to-speech synthesizer to determine which viseme transitions to use, and the rate at which the morphing process should occur. In this manner, we are able to synchronize the visual speech stream with the audio speech stream, and hence give the impression of a videorealistic talking face.

1 Introduction

There has been an increased interest recently in the development of text-to-audiovisual speech synthesis (TTAVS) systems, in which standard text-to-speech (TTS) synthesizers are augmented with a visual component taking on the form of a talking face. This interest is driven by the possible deployment of these systems as visual desktop agents, digital actors, and virtual avatars. In addition, these TTAVS systems may also have potential uses in very low bandwidth videoconferencing and special effects, and would also be of interest to psychologists who wish to study visual speech production and perception.

An important aspect which might be desired of these facial TTAVS systems is *videorealism*: the ability of the final audiovisual output to look and sound exactly as if it were produced by a real human face that was recorded by a videocamera.

Unfortunately, much of the recent work in this field falls short of producing the impression of videorealism. The reason for this, we believe, is that most of the current TTAVS systems (Cohen & Massaro 93, LeGoff & Benoit 96) have chosen to integrate 3D graphics-based facial models with the audio speech synthesis itself. Although it is possible to improve visual realism through texture-mapping techniques, it seems that there is an *inherent* difficulty in modelling both the complex visual appearance of a human face and the underlying facial mouth movement dynamics using 3D graphics-based methods. In this work, therefore, we try to alleviate some of these problems by exploring an altogether different technique of building a videorealistic talking facial model.

2 Overview of Our Approach

Our approach may best be summarized as an *image-based*, *morphing* method:

1. First, a visual corpus of a subject enunciating a set of key words is initially recorded. Each key word is chosen so that it contains one American English viseme. For simplicity, we assume a one-to-one mapping between phonemes and visemes, and ignore coarticulation effects (Cohen & Massaro 93). Consequently, because there are 40-50 American English phonemes (Olive, Greenwood, et al. 93), the subject is asked to enunciate 40-50 words.

2. Next, one single image for each viseme is identified and extracted from the corpus sequence. This is done manually by searching through the recorded frames.

3. Thirdly, we define, in a manner described in more detail below, a *transformation* from each viseme image to every other viseme image. Thus, if there are 40 visemes in our final set, we define 40^2, or 1600, such transformations.

4. Finally, we utilize a text-to-speech system (Black & Taylor 97) to convert unconstrained input text into a string of phonemes, along with duration information for each phoneme. Using this information, we determine the appropriate sequence of viseme transitions to make, as well as the rate of the transformations. The final visual sequence is composed of a *concatenation* of the viseme transitions, played in synchrony with the audio speech signal generated by the TTS system.

3 Morphing

The key aspect to the above approach is being able to easily define and construct realistic transformations between each of the visemes. The approach we have adopted in this work is to use *morphing*.

Morphing was first described by Beier & Neely 92 in the context of generating smooth and realistic transitions between different faces for Michael Jackson's *Black or White* music video. The transformations between images occur as a *warp* of the first image into the second, a similar *inverse warp* of the second image into the first, and a final *cross-dissolve* or *blend* of the warped images. In work that is very similar in spirit to ours, Scott, Kagels, et al. 94, and Watson, Wright, et al., also noticed the viability of using morphing as a method to transition between different mouth viseme imagery.

The difficulty with traditional morphing approaches is that the specification of the warp between the images requires the definition of a set of high-level *features*. These features serve to ensure that the warping process preserves the desired correspondence between the geometric attributes of the objects to be morphed. For example, if we were morphing between two faces, we would want the eyes in one face to map to the eyes in the other face, the mouth in one face to map to the mouth in the other face, and so on. Consequently, the correspondence between these eyes and mouth features would need to be specified.

When it is done by hand, however, this feature specification process can become quite tedious and complicated, especially in cases when a large amount of imagery is involved. In addition, the process of specifying the feature regions usually requires hand-coding a large number of ad-hoc geometric primitives such as line segments, cornerpoints, arcs, circles, and meshes. As a result, we have resorted to *optical flow methods* to alleviate these problems.

4 Optical Flow

Optical flow was originally formulated by Horn & Schunck 81 in the context of measuring the motion of objects in images. This motion is captured as a two-dimensional vector field that describes how each pixel has moved between the images. From our perspective, optical flow is thus important because it allows for the *automatic* determination of the locations of the feature points across images. In this framework, however, there is one feature point for each pixel in an image, which allows us to bypass the need for hand-coding any feature primitives.

One possible approach, therefore, to produce a viseme transformation is to compute optical flow between two viseme images, and then to warp the starting viseme along the computed flow to create various intermediate images. However, this approach usually *fails* to produce realistic transformations, for two reasons:

Firstly, in many cases the optical flow algorithms are not able to compute reasonable motion estimates between viseme imagery. The reason for this is that most algorithms, including ours (Bergen & Hingorani 90), typically make a *motion smoothness assumption*, in which the pixel displacements between images are assumed to be small and locally similar. This assumption is necessary in order to make the problem of determining optical flow well-posed and tractable. In practice, however, mouth pixel displacements between viseme imagery can be significantly large, so the motion smoothness assumption is violated, and the optical flow algorithms cannot obtain a good estimate of the mouth transformation between visemes.

Secondly, even if the computed optical flow vectors are reasonably accurate, the process of warping the first viseme along the optical flow usually does not lead to a realistic transformation into the second viseme. The reason for this is that a large number of novel pixels may have "become visible" in the second viseme image which did not exist in the first viseme. Since the warping process essentially just moves pixels around, this *visibility* phenomenon, which is quite common in viseme imagery due to the opening and closing of the mouth, cannot be modelled sufficiently.

Fortunately, we have found that these shortcomings can be remedied using some simple heuristics.

5 Concatenated Optical Flow

In cases where the computed optical flow between two visemes is not good enough, we have found that a *concatenation* procedure improves the final flow estimates considerably. Since the original visual corpus is digitized at 30 fps, there are many intermediate frames that lie between the chosen viseme frames. The pixel motions between these consecutive frames are small, and hence the motion smoothness assumption is not violated. Consequently, we compute a series of consecutive optical flow vectors between each intermediate image and its predecessor, and then *concatenate* them all into one large flow vector that defines the global transformation between the chosen visemes.

It is not practical, however, to compute concatenated optical flow between viseme images that are very far apart in the recorded visual corpus. The repeated concatenation that would be involved across the hundreds of intermediate frames leads to a considerably degraded final flow. Consequently, we have found that the best procedure for obtaining good correspondences between visemes is actually a *mixture of both direct and concatenated flow computations*: typically, an intermediate frame is chosen that is simultaneously similar in shape to the chosen starting viseme, and also close in distance to the chosen ending viseme. Direct correspondence is then computed between the starting viseme and this intermediate frame, and concatenated flow is computed from the intermediate up to the ending viseme. The final flow from the starting viseme to the ending viseme is then itself a concatenation of both of these direct and concatenated subflows.

Further details of the concatenation procedure itself may be found in Ezzat 96.

6 Morphing Along Optical Flow

The visibility problems discussed previously may be remedied by *morphing* along the optical flow vectors themselves! This was an important observation made initially in Bergen & Hingorani 90, and then subsequently in Beymer, Shashua, Poggio 93. The reason for this is that morphing involves *two* warps, one from the starting viseme to the intermediate point, and another from the ending viseme to the same intermediate point. The two warped images are then scaled by respective *blending parameters* and then added to produce the final morphed image.

The key to overcoming the visibility issues is to *interpolate the blending parameters* so that they initially place greater emphasis on the starting viseme, and then gradually reverse the situation so that greater emphasis is placed on the ending viseme. In this manner, as the morph proceeds, the blending stage "fades out" the warped versions of the starting viseme and "fades in" the warped versions of the ending viseme. The blending process thus allows the two warps to be effectively combined, and the "new" pixels of the second viseme to become involved in the viseme transition itself.

We have tested the above approach, and found it to produce extremely realistic transitions between a wide variety of viseme imagery, including transitions between open and closed mouths. Occasionally, the process does produce peculiarities, but ultimately the rate of transitions between viseme imagery is so fast that these peculiarities do not stand out visibly. Furthermore, the fact that our techniques allow us to produce reasonably accurate viseme transitions in an automatic manner is an added advantage.

For illustration, an example of a typical morph along optical flow vectors between the `\m\` and the `\ah\` viseme is shown in figure 1. Details of our morphing procedure may be found in Ezzat 96.

7 Advantages of Our Model

Our model for a videorealistic talking face may thus be viewed as a collection of viseme imagery and the set of optical flow vectors defining the morph transition paths from every viseme to every other viseme.

There are certain advantages to such a model which should be pointed out:

- First, such a model *reduces* the size of the original visual corpus that needs to be recorded, because we only need to sample the visemes themselves, and not the transitions between them. Instead, our optical flow method allows us to compute reasonable approximations to these transitions in an off-line manner.

- Secondly, the model makes it easy to concatenate the consecutive viseme transitions to produce the final visual speech stream, since the ending viseme of one transition is the same as the starting viseme of the following transition in the stream.

Figure 1: A transition between the `\m\` viseme image (top-left) and the `\ah\` viseme image (bottom right). All other images are morphed intermediates.

- Finally, the representation of a viseme transition as an optical flow vector allows us to morph as many intermediate images as necessary to maintain synchrony with the audio produced by the TTS.

8 Audio Synchronization

We have incorporated the Festival TTS system (Black & Taylor 97), developed at the University of Edinburgh, into our work. A *voice* in the Festival system consists of a set of recorded *diphones*, which are stored as LPC coefficients and corresponding residuals (Hunt, Zwierzynski, et al. 89). It is interesting to note that the final audio speech stream is constructed by concatenating the appropriate diphones together, in a manner that is completely analogous to our method for concatenating viseme transitions.

The Festival TTS system models speech production in the traditional manner as a pitch impulse train that is modulated by a vocal transfer function. This model has been historically important for speech synthesis because it effectively isolates the intonation and duration information, captured by the pitch impulse train, from the phonemic information, captured by the vocal filter.

The TTS system thus takes as input a typed sentence and computes as an intermediate representation the desired pitch train with which to excite the vocal transfer function. For each pitch impulse in the train, the TTS system determines its length in samples, and the diphone filter which it will excite. For example, the pitch train for the word `bed` contains a series of impulses that excite various consecutive portions of the diphone `\b-e\`, followed by a series of impulses that excite various consecutive portions of the diphone `\e-d\`.

The information contained in the pitch impulse train is sufficient for creating a visual speech stream in close synchrony with the audio stream. We place a new viseme image at every pitch impulse which excites a diphone different from the previous. So the `\e\` viseme in the previous example would be placed at the position of the first pitch impulse which transitions between the `\b-e\` and `\e-d\` diphones.

The number of frames to morph is determined by counting the total length in samples of all the pitch impulses between any two viseme images. We divide this sample total by the audio sampling rate (16kHz) to determine the duration of a viseme transition in seconds. Multiplication by the desired video frame rate (60fps) then determines the number of needed frames. It should be noted that, in a manner similar to one taken in Scott, Kagels, et al. 94, diphthongs are modelled using two viseme images, even though they are represented in speech as one phoneme.

We have found that the use of TTS timing and phonemic information in this manner produces superb quality lip synchronization between the audio and the video. The drawback of using a TTS system, however, as opposed to a recorded natural speech signal that is manually annotated, is that the audio may have a slightly 'metallic' quality to it. Nevertheless, the flexibility of having our TTAVS system produce audiovisual output for *any* typed text has offset any misgivings we might have regarding final audio quality. We also believe that future generations of TTS systems will continue to achieve better levels of audio quality.

9 Results

We have constructed several audiovisual sentences to test our overall approach for visual speech synthesis and audio synchronization described above. Our results may be viewed by accessing our World Wide Web home page at `http://cuneus.ai.mit.edu:8000/research/miketalk/miketalk.html`

10 Conclusion

In general, we have found that our techniques achieve a significant improvement in realism over other 3D graphics based methods. Nevertheless, there is still quite a long way to go before we can achieve complete videorealism. In particular, there is a clear need to incorporate into our work higher-level mechanisms in visual speech communication such as coarticulation, eye blinks, eye gaze changes, eyebrow movements, and head nods. Future work will concentrate along these lines.

It should be noted that other recent work in this field that uses a different approach was done by Bregler, Covell, et al. 97.

11 References

T. Beier and S. Neely. (1992) Feature-based image metamorphosis. In *SIGGRAPH '92 Proceedings*, pages 35–42, Chicago, IL.

J.R. Bergen and R. Hingorani. (1990) Hierarchical motion-based frame rate conversion. Technical report, David Sarnoff Research Center.

D. Beymer, A. Shashua, and T. Poggio. (1993) Example based image analysis and synthesis. Technical Report 1431, MIT AI Lab.

A. Black and P. Taylor. (1997) *The Festival Speech Synthesis System*. University of Edinburgh.

C. Bregler, M. Covell, and M. Slaney. (1997) Video rewrite: Driving visual speech with audio. In *SIGGRAPH '97 Proceedings*, Los Angeles, CA.

M. M. Cohen and D. W. Massaro. (1993) Modeling coarticulation in synthetic visual speech. In N. M. Thalmann and D. Thalmann, editors, *Models and Techniques in Computer Animation*, pages 139–156. Springer-Verlag, Tokyo.

T. Ezzat. (1996) Example-based analysis and synthesis for images of human faces. Master's thesis, Massachusetts Institute of Technology.

B. K. P. Horn and B. G. Schunck. (1981) Determining optical flow. *Artificial Intelligence*, 17:185–203.

M. J. Hunt, D. A. Zwierzynski, and R. Carr. (1989) Issues in high quality lpc analysis and synthesis. In *Proceedings of EUROSPEECH '89*, volume 2, pages 348–351, Paris, France.

B. LeGoff and C. Benoit. (1996) A text-to-audiovisual-speech synthesizer for french. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP '96)*, Philadelphia, USA.

J. Olive, A. Greenwood, and J. Coleman. (1993) *Acoustics of American English Speech: A Dynamic Approach*. Springer-Verlag, New York, USA.

K.C. Scott, D.S. Kagels, S.H. Watson, H. Rom, J.R. Wright, M. Lee, and K.J. Hussey. (1994) Synthesis of speaker facial movement to match selected speech sequences. In *Proceedings of the Fifth Australian Conference on Speech Science and Technology*, volume 2, pages 620–625.

S.H. Watson, J.R. Wright, K.C. Scott, D.S. Kagels, D. Freda, and K.J. Hussey. An advanced morphing algorithm for interpolating phoneme images to simulate speech. Unpublished technical report, Jet Propulsion Laboratory, California Institute of Technology.