TalkingFace: Using Facial Feature Detection and Image Transformations for Visual Speech

Ali Arya , Babak Hamidzadeh

Dept. of Electrical & Computer Engineering, University of British Columbia,
2356 Main Mall, Vancouver, BC, Canada V6T 1Z4, Phone: (604)822-9181, Fax: (604)822-5949
Email: {alia , babak}@ece.ubc.ca

ABSTRACT

Visual presentation of a talking person requires the generation of image frames showing the speaker in various views while pronouncing various phonemes. The existing approaches, mostly use either a complex 3D geometric model to reconstruct a desired image or a set of 2D images for each viewpoint, to select from. We propose a new system which utilizes facial feature detection and image-based transformation to create any talking frame using only one given image from desired viewpoint and a set of reference images from one standard view. The proposed approach, together with optical flow-based view morphing and a customizable concatenative Text-To-Speech, makes a personalized visual speech generation system which can be used for moving/talking head applications where an optimal trade-of between computational complexity and image database requirements is necessary.

1. INTRODUCTION

The recent developments of interactive multimedia systems have raised an ever-growing demand for software agents. These agents need to mimic the audio and visual aspects of a human being in a virtual world. The main advantage of such agents is the ability to create a multimedia presentation using a limited pre-recorded database and a set of control commands instead of on-demand multimedia recording and transfer which might be hard or even impossible due to limited storage or data transfer bandwidth, or unavailability of the characters. Videoconferencing, training and technical/customer support, and visual effects in the movies are just some sample applications of these personalized agents.

Computer animation researchers have been working on developing 3D models of human head and defining its basic operations for a long time [2,3,6,7]. There has also been a considerable interest in multiple view 2D models [1,4,5]. We believe that, as observed in many experiments, pure image-based approaches, without considering geometrical features and 3D aspects, need a considerably huge database and/or fail to create valid facial presentations. On the other hand, traditional 3D models usually need complex computations and even hardware (e.g. laser range finders). We propose a hybrid approach which is basically image-based but utilizes facial feature detection to create talking images. Our approach uses a limited set of images to define necessary image transformations for talking or simple head movements, and then given any arbitrary view of the same or even a different person, it applies those transformations to create moving/talking images in the desired views.

In the next section, we briefly review some related works in facial animation. In Section 3, our basic image-based approach to Text-To-Visual-Speech is described. The feature-based improvement to this basic system (in order to handle multiple views) is the subject of Section 4. Some experimental results and concluding remarks will be presented in Sections 5 and 6.

2. RELATED WORKS

The existing approaches to moving/talking heads use either image-based 2D models [1,4,5] or geometry-based 3D ones [2,6,7]. 3D models provide more power to construct the head in any given view, but they are hard to build and usually lack the realistic appearance even after texture mapping. Recently, most of these models tend to follow the MPEG-4 standard Face Definition/Animation Parameters [4] which resembles Facial Action Coding System, FACS [3].

Image-based approaches, use a multiple-view model of the head and create a desired view by applying some transformations (e.g. linear combination) to a set of standard 2D views. Complexity of head structure, non-rigid motion, lack of one-to-one correspondence between pixels (new and overlapping points), and solving the correspondence problem when mapping points to each other, are among the issues in this approach. Ezzat, et al [4], use view morphing between prerecorded visemes (facial views when pronouncing different phonemes) to create a video corresponding to any speech. Bregler , et al [1], combine a new image with parts of existing footage (mouth and jaw) to create new talking views.

Both these approaches are limited to a certain view where the recordings have been made. No transformation is proposed to make a talking view after some new movements of the head. In a more recent work based on [4], Graf, et al [5], propose recording of all visemes in a range of possible views, so after detecting the view (pose) proper visemes will be used. This way talking heads in different views can be animated but the method requires a considerably large database.

By detecting the facial features and determining the transformations that they go through to create a certain head movement (including talking), it is possible to apply those movements to a new image, even of a new person. This way, we neither need a complex geometrical 3D model nor a large database of images. This idea forms the basic concept of our proposed approach as discussed in next sections.

3. PERSONALIZED MOVING AND FIXED-VIEW TALKING HEAD

A concatenative Text-To-Speech (TTS) engine and a view morphing-based frame generator are the building blocks of our basis system, as shown in Figure 1. This system can generate multimedia output for simple head movements and also talking in a fixed view. The video subsystem uses an image-based approach, i.e. we do not utilize geometric models of head/face which are usually hard to calculate. Instead, we base our approach on a set of input images of the head in some key positions and also in visemes in frontal view (Figure 2, a-c). We define simple movements as a transition from one of these views to another, assuming that it is either pure talking or pure moving. Such a transition is created by a morphing process, i.e. applying a correlation-based optical flow algorithm to source and destination views, finding a flow vector, and then applying that vector to the source image incrementally, to create as many intermediate images as we need. Correlation-based optical flow algorithms are more powerful compared to gradient-based ones, in solving the correspondence problem required for morphing, specially for larger movements. To enhance the images, and to handle the holes and overlaps, we perform a forward and a backward warp similar to [4], and build the final morph by taking the average of forward and backward images.

The major issue here is the optical flow matching error. We minimize this error by applying a hierarchical correlation-based optical flow algorithm (also useful in speeding up the calculations) which is more capable of coping with geometric moves, but the matching error can still be more than ten percent for large movements of the head. This causes visual noise-like pixels in the final image.

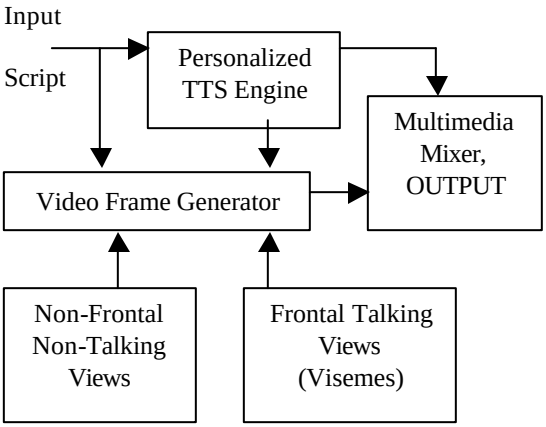

Figure 1. Basic Block Diagram

The second part of our system is responsible for making the audio stream. The input text to be spoken is sent to a TTS engine which is capable of providing a sequence of phonemes rather than actual audio output. Based on this phoneme sequence, a set of pre-recorded diphones are selected, concatenated, and eventually mixed with the video stream. The initial diphone database can be manually created or automatically extracted from a given piece of speech. The diphones are pre-scaled for power and pitch, and will be concatenated with a degree of overlap to minimize the discontinuity. We use spectral distance to dynamically find the best point to connect two diphones, and also to extract diphones from a given audio. The dynamic length of diphones (original length minus the overlap) determines the number of frames to be generated in order to guarantee the lip-synchronization.

4. FACIAL FEATURE-BASED VISUAL SPEECH

4.1. Combined Movements

Another type of movement is moving the head while talking, i.e. talking in a non-frontal view. Our main contribution in this approach is to provide a simple yet effective method to convert a non-talking arbitrary image from any viewpoint to a talking one corresponding to any visemes (or phoneme). This enables us to have visemes in any view, so make talking heads in non-frontal views and/or motion.

Assuming I_1, I_2, and I_3 are non-talking frontal, non-talking non-frontal, and talking frontal views, the objective is to create I_4 which is a non-frontal talking view. To perform this, we make a new mapping vector V_{24} (V_{mn} maps from I_m to I_n) by combining V_{21} and V_{13}. For each pixel in I_2, we find the corresponding pixel in I_1 using the backward map V_{21}, then we use the corresponding value from V_{13} as the new mapping vector of original pixel in I_2. The resulting map will

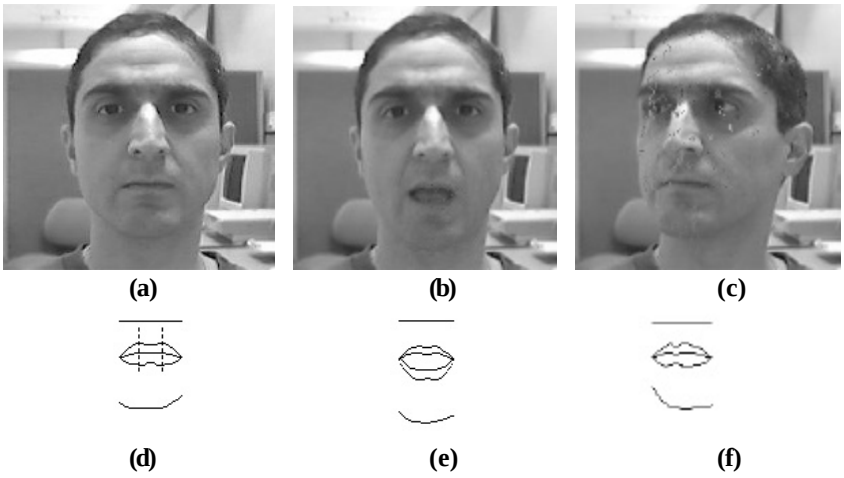

Figure 2 (a) sample head images, (b-c) reconstructed talking and moving images, (d-f) the facial features. (d) shows how the specially detected points on the upper lip can define three regions in the feature area, used later when creating transformed images.

cause the points in non-frontal view to have the same transition as their corresponding points in frontal view when talking. The holes which will appear after this process can be filled up, approximately, using pixels in I_3 after being transformed using V_{12}.

In case of combined movements, the error mentioned in simple movements shows itself in a more serious way due to additive nature of the transformation. This results in less-than-ideal quality of the final image for large head movements and demonstrates the inability of pure image-based approach (without using any domain information) in handling complex moves. Based on this observation, we have implemented a hybrid system as our second approach, which uses some 2D geometric information about the subject (face) to enhance the flow maps between existing images and also create new feature information for the images to be created for combined movements.

4.2. Facial Feature Detection

As the first step, a facial feature detection algorithm will be applied to any image which detects the anchor points. We use threshold-based methods and also color comparison to detect regions and extract a feature matrix which consists of following feature vectors:

1- Horizontal line at the bottom of nose
2- Lip4 to Lip1 for upper and lower lips
3- Jaw

The vectors have the same size which is determined by the distance between right and left corner of the lips. Figure 2 illustrates sample head images with their detected features. In addition to the lip corners, two distinctive points in the upper lip are also detected, as shown in the Figure 2. The following parts of our approach assume that these information are available by the feature detection module, i.e. the required features are not hidden due to head movements.

4.3. Feature-based Flow Map Correction

Inside the region specified by the feature matrix (nose to jaw, right to left lip corners) the flow map computed using correlation-based optical flow will be modified by a feature correspondence algorithm. This algorithm uses maximum correlation to find corresponding points along the feature lines, and then interpolates the mapping values between each two feature lines to find the mapping vector for other points.

The following psudu-code describes this algorithm:

```
for column=0 to w//feature matrix width
    for i=0 to 5     //number of feature lines
        //fp1 and fp2 are points on feature lines
        fp2[i] = corr(fp1[i], fline2[i])
    interpolate()//other points
```

4.4. Combined Facial Features

The combined movements as described before, involve creation of a completely new image rather than interpolating between two existing ones. To perform this, we follow the method explained in 4.1, but we add the facial feature-based correction. The overall procedure to create a new, e.g. non-frontal talking view, image is as follows:

1- create feature matrix for three existing images
2- apply feature-based corrections to flow maps
3- combine existing feature matrices to create target feature matrix
4- apply the algorithm in 4.1 to make a new map
5- fine-tune the map using the newly created feature matrix

(a) **(b)** **(c)**

Figure 3. Sample results of creating new (a) talking image and (b) its feature matrix, using (c) existing talking non-frontal view. It should be noted that the mapping vectors for lip movements are extracted from a different set of images (Figure 2) which can be from a totally different person.

To create a new feature matrix, we basically follow the same concepts as for maps, i.e. find the corresponding points and apply the same transformation. Corresponding feature points are found with a correlation-based search limited to the points on the related feature line. In cases where two images are not simply a "moved" version of each other (e.g. when applying the speech to a new person's image), we use two specially detected feature points on the upper lip to divide the feature area to three regions, and in each region we map the points linearly (see Figure 2).

Finally, the region inside the mouth in the new image is filled with a transformed version of the same region in the frontal view. This transformation assumes that the points inside the mouth follow the same mapping as the middle portion of the upper lip.

5. EXPERIMENTAL RESULTS

The experimental results show the effectiveness of our approaches in creating an audio/visual presentation. The main advantage of the first (correlation-based optical flow) approach is its simplicity, but as observed in the results (Figure 2), the quality is not satisfactory specially in cases where head has relatively large movements. Figure 3-a shows example results of feature-based transforms used to create a talking view. This approach is more general and powerful and gives better results but it needs more computation for feature detection.

6. CONCLUSION

The main advantage of our combined image-based and feature-based approach is increased precision and reliability in calculating the flow vectors without any need to have a large database of images, and also the ability to use the flow vectors related to the talking of one person, in order to create a talking view of another person. Since we do not perform any correlation-based matching between two talking and non-talking images, these two do not need to be of the same person. This means that an available actor can be used to make high quality frontal views of a talking person which are used to extract the related flow vectors, off line. These vectors can then be applied to any person's images (after feature extraction) to create new talking views.

In future, this approach can be extended to include facial expressions and other head movements. In this way, for any particular facial change, a flow map (image transformation) can be found using a correlation-based optical flow algorithm improved by feature detection and based on a reference person. For any given personal image, this transformation can be applied after feature detection and proper normalization.

7. REFERENCES

[1] C. Bregler, et al, "Video Rewrite," *ACM Computer Graphics*, 1997

[2] V. Blanz and T. Vetter, "A Morphable Model For The Synthesis Of 3D Faces," *Proc ACM SIGGRAPH*, 1999.

[3] P. Ekman and W. V. Friesen, *Facial Action Coding System*, Consulting Psychologists Press Inc., 1978.

[4] T. Ezzat and T. Poggio, "MikeTalk: A Talking Facial Display Based on Morphing Visemes," *Proc IEEE Conf Computer Animation*, 1998.

[5] H. P. Graf, et al, "Face Analysis for the Synthesis of Photo-Realistic Talking Heads," *Proc IEEE Conf Automatic Face and Gesture Recognition*, 2000.

[6] W. S. Lee, et al, "MPEG-4 Compatible Faces from Orthogonal Photos," *Proc IEEE Conf Computer Animation*, 1999.

[7] F. Pighin, et al, "Synthesizing Realistic Facial Expressions from Photographs," *Proc ACM SIGGRAPH*, 1998.