A PERSONALIZED TALKING/MOVING HEAD PRESENTATION USING IMAGE-BASED TRANSFORMATIONS

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

This paper addresses the problem of multimedia presentation of a moving and talking head. Existing approaches either use a 3D geometrical representation or multiple 2D views to model and reconstruct faces. To achieve realistic appearance and personalization, and avoid complex models, we propose two approaches. First, we show how pure image-based view morphing can be used to create new moving and talking faces based on existing footage. These views form a video stream which in turn will be synchronized with the output of a personalized Text-To-Speech system. Then, to improve this approach, we use facial feature detection to fine tune the correlation-based optical flow computation used for view morphing, and also create talking views of a new person based on one non-talking view in any position and talking views of a reference person.

1. INTRODUCTION

Talking heads can play a very important role in applications like video conferencing, training, and entertainment. The major drawbacks of existing systems are lack of realistic appearance and personalized audio/visual features, and limited range of movements (e.g. talking only in frontal view). An ideal talking head should be able to replace a person with a virtual agent in an interactive multimedia environment, allowing different movements and speeches requested by an on-line user (through mouse movements, arrow keys, or any other input mechanism) while mimicking the person’s voice and facial appearance.

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]. Recently, following the trend in object recognition, there has a 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, 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 uses facial feature detection to provide a simple yet effective talking/moving head presentation system.

In the next section, we briefly review some related works in facial animation. In Section 3, our basic image-based approach is described. The feature-based improvements to this basic system is the subject of Section 4. The audio subsystem will be reviewed in Section 5, and some experimental results and concluding remarks will be presented in Sections 6 and 7.

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. IMAGE-BASED TALKING HEAD

The system described in this paper consists of two parts, a video and an audio subsystems, which make a fully personalized multimedia presentation. Figure 1 shows the basic block diagram of this system. The video subsystem is responsible for creating a stream of video frames corresponding to a certain movement. We propose two different but related approaches in this regard.

The first approach is image-based, i.e. we do not use 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. 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.

Another type of movement is moving the head while talking, i.e. talking in a non-frontal view. 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 cause the points in non-frontal view to have the same transition as their corresponding points in frontal view when talking.

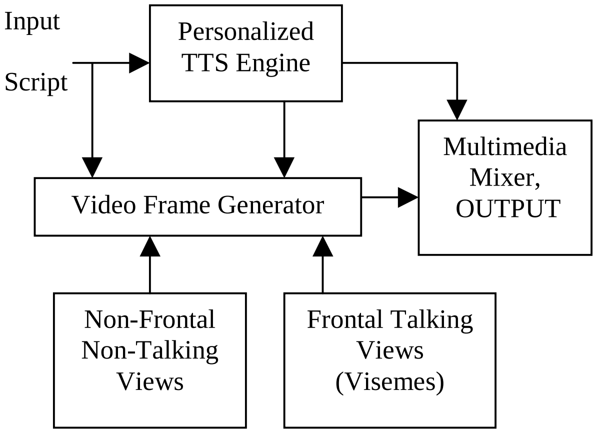


Figure 1. Basic Block Diagram

The holes which will appear after this process can be filled up, approximately, using pixels in I_3 after being transformed using V_{12}. Our main contribution in this approach is to provide a simple yet effective method to create talking heads in any given view.

The second part of our system is responsible for making the audio stream. The input text to be spoken is sent to a Text-To-Speech (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. FEATURE-BASED IMPROVEMENTS

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 fine tune (1) the correspondence matching, and (2) the combined transformation.

As the first step, a facial feature detection algorithm will be applied to any image which detects the anchor points (e.g. lips corners and jaw line with or without manually set markers). These points will be used as the initial values for the optical flow algorithm in finding the corresponding points, i.e. points around them are to be mapped to some points around the corresponding anchor in the second image. After finding a set of more reliable flow vectors for simple movements, we use them in the combined ones (i.e. talking in non-frontal views).

This can be done in two different ways. We can follow the method used in the first approach, apply the backward map, find the corresponding points in frontal view and their vector for talking, and then apply them to the points in non-frontal view. The other way, is to use only the flow vectors for anchor points and interpolate the other vectors based on these ones, in a region around mouth. The first method will provide a more realistic image (possible non-linear changes between the anchor points) but due to existing matching errors, can suffer from mismatched points that show themselves as visual noise. On the other hand, this method does not depend completely on the feature detection stage which itself is error-prone. The second method will result in a noise-free image which may look a little artificial and depends heavily on the correctness of feature detection. Combination of both of these methods seems to provide the best results. This can be done by simply performing a weighted average, or using the interpolated vectors as a rule base for accepting/modifying the other group of flow vectors.

5. AUDIO SUBSYSTEM

As shown in Figure 1, the video subsystem receives input from main input script to handle head movements and also the audio subsystem for a list of phonemes to be translated to visemes. We use a standard TTS engine to convert the input text to a sequence of phonemes. This data is then used in two different ways:

a) As input to the video subsystem
b) In the audio subsystem to make a personalized sound

The personalized sound is made by concatenating diphones. A corpse of diphones is extracted from a given speech of the corresponding person using a Hidden Markov Model, used to automatically segment a given speech into diphones. The smooth connection of diphones also needs special considerations and can not be done by simply appending an audio segment to the end of another one. Three types of problems can happen when concatenating diphones:

a) Power level difference
b) Pitch mismatch
c) Diphone-boundary discontinuity

The first two issues are usually related to differences at the time of recording which are very important when the input audio is from different sources. We have used power equalization and pitch scaling algorithms to handle these differences [8].

The third issue arises due to either non-ideal creation of diphones (not starting and ending at the middle of phoneme steady states) or audio difference in transition from a phoneme to previous and next phonemes. These both result in discontinuity of speech at the boundary of diphones. Two time-domain (based on signal correlation) and frequency-domain (based on comparing FFT of signals) are applied to find the best connection point for two diphones. The frequency-domain approach seems to be more adequate and reliable. To allow the best performance for this algorithm, we create the diphones so that the length is a little longer than mid-phoneme to mid-phoneme.

5. EXPERIMENTAL RESULTS

The experimental results show the effectiveness of our approaches in creating an audio/visual presentation. The main advantage of the first approach is its simplicity, but as observed in the results, the quality is not satisfactory in all cases. Figure 2 shows example results of pure image-based transforms used to create a talking view.

The second approach is more general and powerful and gives better results but it needs more computation for feature detection. Some experimental results are demonstrated in Figure 3 which compares the image-based and feature-based approaches.

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: Driving Visual Speech with Audio," *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.

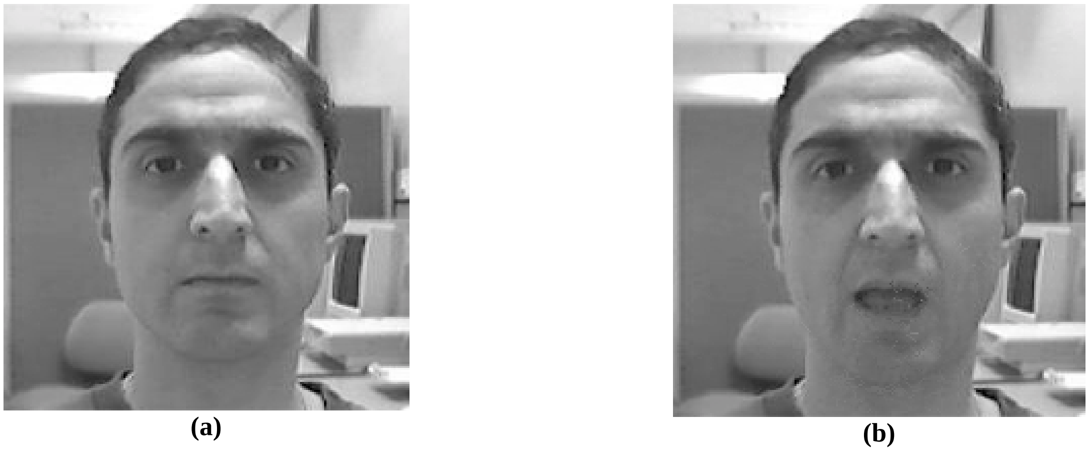
[4] T. Ezzat and T. Poggio, "MikeTalk: A Talking Facial Display Based on Morphing Visemes," *Proc IEEE Conf Computer Animation*, 1998.

[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.

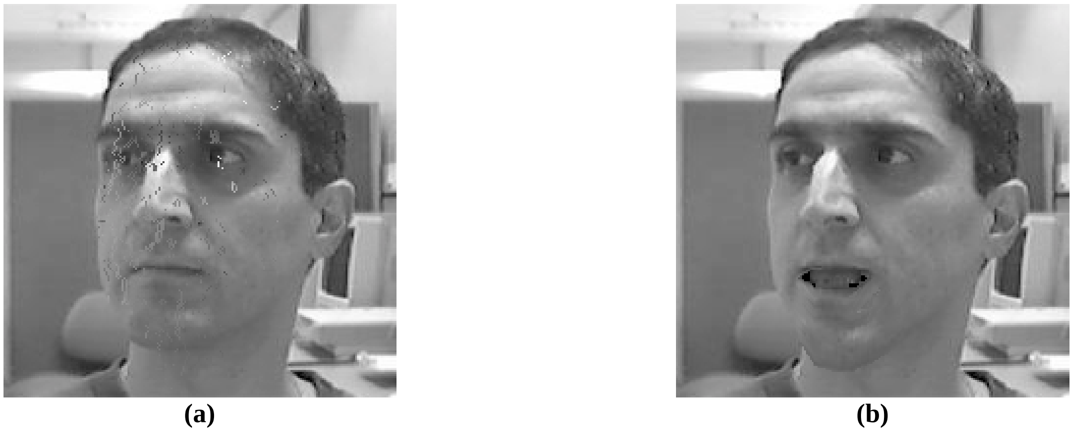
[7] F. Pighin, et al, "Synthesizing Realistic Facial Expressions from Photographs," *Proc ACM SIGGRAPH*, 1998.

[8] L. R. Rabiner, and R. W. Schaffer, *Digital Processing of Speech Signal*, Prentice-Hall, 1978



(a) (b)

Figure 2. Image-based Mapping from Non-Talking Frontal View (a) to Talking (b)



(a) (b)

Figure 3. Image-based Mapping from Non-Talking Frontal View (2-a) to Non-Talking Non-Frontal (a), and Non-Frontal Talking View (b) created using Feature-based Approach