Image-based Talking Head: Analysis and Synthesis

Kang Liu[1], Joern Ostermann[2]

[1] *Institut für Informationsverarbeitung, Leibniz Universität Hannover, Email: kang@tnt.uni-hannover.de*

[2] *Institut für Informationsverarbeitung, Leibniz Universität Hannover, Email: ostermann@tnt.uni-hannover.de*

Abstract

In this paper, our image-based talking head system is presented, which includes two parts: analysis and synthesis. In the analysis part, a subject reading a predefined corpus is recorded first. The recorded audio-visual data is analyzed in order to create a database containing a large number of normalized mouth images and their related information. The synthesis part generates natural looking talking heads from phonetic transcripts by the unit selection algorithm. The phonetic transcripts can be extracted from a TTS (Text-To-Speech) system (for text-driven animation) or from speech by an aligner (for speech-driven animation). The unit selection is to select and concatenate appropriate mouth images from the database by minimizing two costs: lip synchronization and smoothness. The lip synchronization measures how well the unit fits to the phonetic context, and the smoothness cost measures how well two units join together. Finally, the mouth images are stitched at the correct position on the face of a recorded video sequence and the talking head is displayed.

Introduction

The development of modern human-computer interfaces and their applications such as E-Learning and web-based information services has been the focus of the computer graphics community in recent years. Image-based approaches for animating faces have achieved realistic talking heads [1]. Recent investigations [1, 2, 3] indicate that image-based talking heads look more realistic than 3D model-based talking heads, especially when the mouth is open.

An Image-based Talking Head

Our talking head system [3] is divided into two parts: off-line analysis and on-line synthesis.

Analysis

In our studio, a native speaker is recorded while reading a corpus of 150 sentences. These sentences are designed to find a trade off between the English phoneme coverage and the size of the corpus. A lighting system is developed for the audio-visual recording, which minimizes the shadow on the face of an subject and hence reduces the change of illumination on the moving head.

After recording, the audio data and the texts are aligned by an aligner, which segments the phonemes of the audio data. The aligner is first trained for specific voices in order to increase the accuracy of the alignment given the recorded audio and the spoken text. Once the aligner has been trained, the aligner produces a timed sequence of phonemes. Therefore, for each frame of the recorded video, the corresponding phoneme and its phoneme context are known. The phoneme context is required in order to capture coarticulation effects.

Using a 3D head scan of the recorded speaker, the recorded videos are processed by model-based motion estimation [4], which estimates the head motion parameters for each frame. The 3D head scan is a 3D face representation, which is a polygon mesh consisting of a collection of vertices and polygons that define the shape of a face in 3D. The model-based estimation algorithm stores texture information of the object and tries to find the motion parameters of the rigid head in a new frame by minimizing the difference between the textured 3D model and the face in the new frame. Using the motion parameters, the mouth images are normalized. Normalized mouth images are transformed into the PCA space, so that few parameters are needed to describe the appearance of the mouth image. The shape of the mouth images are extracted by AAM (Active Appearance Models), from which geometric parameters such as mouth width and height are derived [5].

The off-line analysis provides a database containing a large number of mouth images. Each image is characterized by geometric parameters (such as mouth width and height), texture parameters (PCA parameters), phonetic context, original sequence and frame number.

Synthesis

Based on the database, the on-line visual text-to-speech synthesis animates a realistic talking head using the unit selection as shown in Fig. 1. The face is divided into a set of facial parts, such as eye and mouth. Each facial part is made of a 3D wireframe model that describes its 3D shape and the database of mouth images that describes its appearance variance. The TTS (Text-To-Speech) synthesizer converts the input text to speech as well as the sequence of phonemes and their durations, which are sent to the unit selection. Depending on the phonetic information, the unit selection selects mouth images from the database and concatenates them in an optimal way. Thereafter, in the image rendering module, the mouth images are first wrapped onto a personalized 3D wireframe model and rotated and translated to the correct position given by an image of the background video. Finally the facial animation is synchronized with the audio, and a talking head is displayed.

The unit selection selects the mouth images correspond-

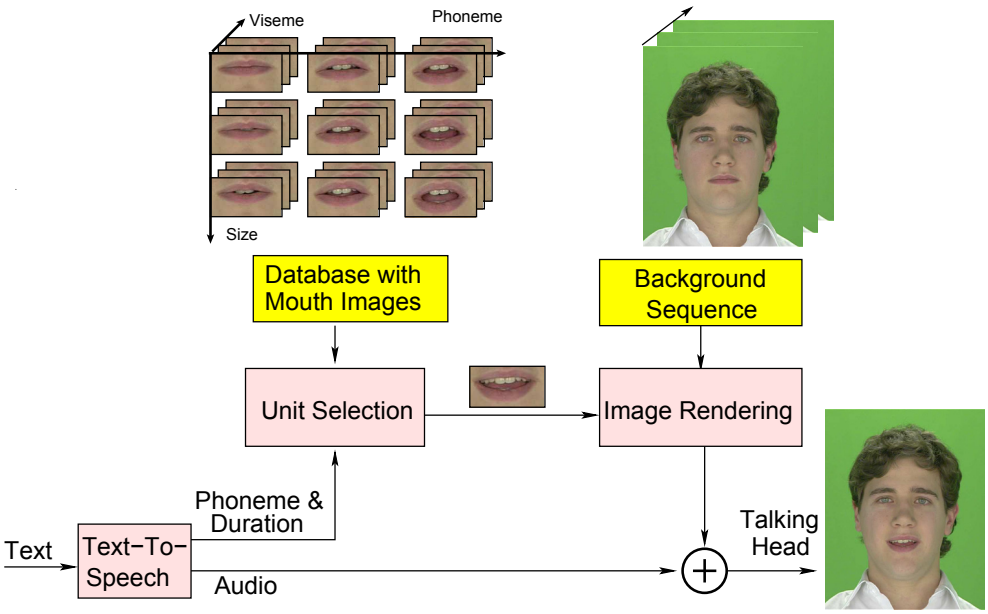

Abbildung 1: System overview of the talking head system.

ing to the phoneme sequence, using a target cost and a concatenation cost function to balance two competing goals: lip synchronization and smoothness of the transition between consecutive images. The cost of lip synchronization considers the co-articulation effects by matching the distance between the phonetic context of the synthesized sequence and the phonetic context of the mouth image in the database. The cost of smoothness reduces the visual distance at the transition of images in the final animation, favoring transitions between consecutively recorded images. A Viterbi search is used to find the optimal mouth image sequence with minimal weighted target and concatenation costs.

The weights for unit selection are trained by Pareto optimization [3], which searches optimal weight sets in the weight space efficiently and optimize several objective measurements.

Experimental Results

Formal subjective tests show that synthesized animations generated by the optimized talking head system matches the corresponding audio naturally. More encouraging, 3 out of 5 synthesized animations are so realistic that the viewers cannot distinguish them from original videos [3]. These animations are available at http://www.tnt.uni-hannover.de/project/facialanimation/demo.

Conclusions

In this paper, we have described an image-based talking head system, which consists of an off-line audio-visual analysis and an on-line unit selection synthesis.

Compared to the reference system [2], our image-based talking head is improved in the three main aspects: In the analysis, head motion is estimated precisely and robustly by using model-based approach rather than by using feature-based approach. Instead of template matching based feature detection, we use AAM based facial feature detection, which is insensitive to illumination changes on the face resulted from head and mouth motions. The unit selection is optimized by Pareto optimization approach.

Acknowledgements

This work is funded by German Research Association (Deutsche Forschungsgemeinschaft - DFG Sachebeihilfe OS295/3-1). This work has been partially supported by EC within FP6 under Grant 511568 with the acronym 3DTV.

Literatur

[1] Fagel, S., Bailly, G., Theobald, B.: Animating Virtual Speakers or Singers from Audio: Lip-Synching Facial Animation. In EURASIP Journal on Audio, Speech, and Music Processing, 2010

[2] Cosatto, E., Ostermann, J., Graf, H., Schroeter, J.: Lifelike talking faces for interactive services. Proceedings of the IEEE, vol. 91, 1406-1429, 2003

[3] Liu, K., Ostermann, J.: Optimization of an image-based talking head system. Special issue on animating virtual speakers or singers from audio: Lip-synching facial animation, EURASIP Journal on Audio, Speech, and Music Processing, 2009

[4] Weissenfeld, A., Urfalioglu, O., Liu, K., Ostermann, J.: Robust Rigid Head Motion Estimation based on Differential Evolution. Proc. ICME 06, 225-228, 2006

[5] Liu, K., Weissenfeld, A., Ostermann, J., Luo, X.: Robust AAM Building for Morphing in an Image-based Facial Animation System. Proc. ICME 08, 933-936, 2008