An Image-based Talking Head System

Kang Liu and Joern Ostermann, Fellow, IEEE

Institut für Informationsverarbeitung, Leibniz Universität Hannover
Appelstr. 9A, 30167 Hannover, Germany
kang@tnt.uni-hannover.de, ostermann@tnt.uni-hannover.de

Abstract

This paper presents an image-based talking head system, which includes two parts: analysis and synthesis. The analysis is to create a database containing a large number of mouth images and their associated facial and speech features. The synthesis is to generate realistic facial animations from phonetic transcripts of text. The facial animation is produced by selecting and concatenating appropriate mouth images that match the spoken words of the talking head. Subjective tests show that 60% of the animations are indistinguishable from real recordings.

Index Terms: talking head, unit selection, evaluation

1 Introduction

Faces are the focus of attention for any audience, and the slightest deviation from normal behaviour is immediately noticed, especially in the region of the mouth. Believable talking faces are essential for applications in face-to-face communication, such as virtual agents of e-commerce and multimedia service. Recent investigations [1, 2] indicate that image-based talking heads look more realistic than 3D model-based talking heads, especially when the mouth is open.

2 System Overview

Our talking head system [2] is divided into two parts: off-line analysis and on-line synthesis. The off-line analysis provides a database containing a large number of mouth images. Each image is parameterized by facial features (such as mouth geometric and texture features), speech features (such as phonetic context), etc.

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 corresponding 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.

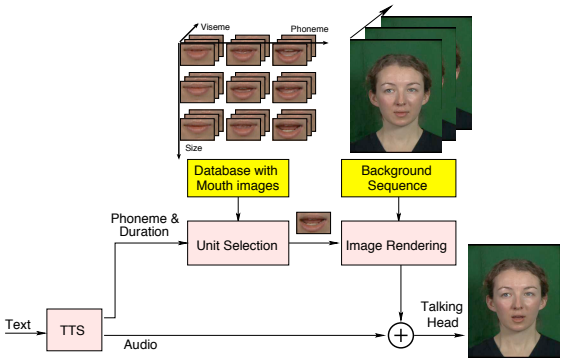

Figure 1: System overview of the talking head system.

The weights for unit selection are trained by Pareto optimization [2], which searches optimal weight sets in the weight space efficiently and optimize several objective measurements.

3 Results

For subjective evaluation, the real and animated video pairs are presented only once one after the other. The viewers are asked to decide whether it is real or animated. The results show that 60% of the animations are indistinguishable from the real videos. The animations for subjective tests are available at http://www.tnt.uni-hannover.de/project/facial...animation/demo/subtest.

References

[1] E. Cosatto, J. Ostermann, H. Graf, and J. Schroeter, "Lifelike talking faces for interactive services," *Proceedings of the IEEE*, vol. 91, pp. 1406–1429, 2003.

[2] K. Liu and J. Ostermann, "Realistic facial animation system for interactive services," in *Proceedings of Interspeech 2008*, 2008, pp. 2330–2333.