Adaptive Man-Machine Interfaces
MIT9904-15

Proposal for 1999-2000 Funding

Tomaso Poggio

Project Overview

We propose two significant extensions of our recent work on developing a text-to-visual-speech (TTVS) system {Ezzat98}. Our existing *synthesis* module may be trained to generate image sequences of a real human face synchronized to a text-to-speech system, starting from just a few real images of the person to be simulated. We propose to 1) extend the system to use morphing of **3D models** of faces -- rather than face images -- and to output a 3D model of a speaking face and 2) **enrich the context** of each viseme to deal with coarticulation issues. The main applications of this work are for virtual actors and Hollywood and for very-low-bandwidth video communication. In addition, the project may contribute to the development of a new generation of computer interfaces more user friendly than today's interfaces.

A Trainable System for Real Time Synthesis of Visual Speech

Overview The goal of our recent work {Ezzat98} bas been to develop a text-to-audiovisual speech synthesizer called MikeTalk. MikeTalk is similar to a standard text-to-speech synthesizer in that it converts text into an audio speech stream. However, MikeTalk also produces an accompanying visual stream composed of a talking face enunciating that text. An overview of our system is shown in Figure 1.

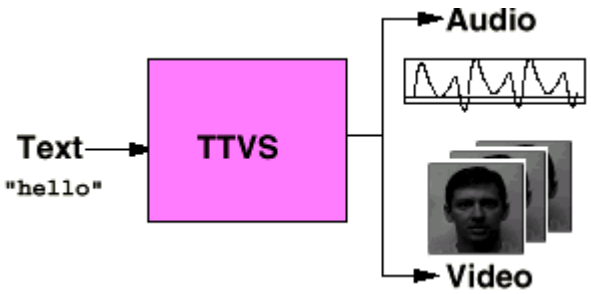

Figure 1.
Input-output behavior of the TTVS module

Text-to-visual (TTVS) speech synthesis systems are attracting an increased mount of interest in the recent years, and this interest is driven by the possible deployment of these systems as visual desktop agents, digital actors, and virtual avatars. In addition, they may also have potential uses in special effects, very low bit rate coding schemes (MPEG4), and would also be of interest to psychologists who wish to study visual speech production and perception. In this work, we are particularly interested in building a TTVS system where the facial animation is *photo realistic*: that is, we desire our talking facial model to look as much as possible as if it were a video camera recording of a human subject, and not that of a cartoon-like human character.

In addition, we choose to focus our efforts on the issues related to the synthesis of the visual speech stream, and not on audio synthesis. For the task of converting text to audio, we have incorporated into our work the Festival speech synthesis system, which was developed by Alan Black, Paul Taylor, and colleagues at the University of Edinburgh \cite{Festival97}. Festival is freely downloadable for non-commercial purposes, and is written in a modular and extensible fashion, which allows us to experiment with various facial animation algorithms.

All prior work (not described for space reasons) involved the incorporation of a facial display based on a parametrized three-dimensional model of some sort, whether it is a simple polygonal model, or a more complex one involving muscles and bone tissue. While they have all certainly achieved notably good results, most of the work has thus far failed to (a) achieve a high degree of video realism in the final output, and (b) result in a model that can be used to build other facial talking models in an easy, automatic manner. In order to address these issues, we would like to explore a different approach that does not resort to any three-dimensional modeling of the human face. Our approach {Ezzat98} is directly motivated by the work of {BeyShaPog93} as well as {Ezzat96}, and may best be summarized as an *image-based, learning* method:

- 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*. A viseme is a recently-coined term for a mouth shape that is associated with spoken speech {BenLal92}. For simplicity, we assume a one-to-one mapping between phonemes and visemes and ignore coarticulation effects {CohenMassaro93}. Consequently, because there are 40-50 American English phonemes {Olive93}, the subject is asked to enunciate 40-50 words.

- Next, one single image for each viseme is identified and extracted from the corpus sequence. This will be done initially by hand by searching through the recorded frames, but ultimately we would like to develop automatic learning-based techniques to extract these visemes, possibly using the analysis module.
- Thirdly, we construct, a *morph transformation* from each viseme image to every other viseme image. The morph transformation is constructed in two substeps: first, optical flow {HorSch81} is computed between the two visemes, obtaining a dense vector field representing the motion of the mouth between both images. The second is to morph {BeiNee92} the two visemes along the optical flow vectors themselves. An example of a typical morph along optical flow vectors between visemes is shown in figure 2.

Figure 2.
A transition between the \m\ and the \ah\ viseme image (bottom right). All other images are morphed intermediates

- Finally, we utilize a text-to-speech system {SproatOlive95, Festival97}) 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.

Future goals

The key future goals that we plan to attain with the NTT funding are:

1) enrich our approach with context to deal with the coarticulation problem. While our earlier work {Ezzat98} made strong strides towards photorealism, it did not address the *dynamic* aspects of mouth motion. Modelling dynamics requires addressing a phenomenon termed *coarticulation* {CohenMassaro93}, in which the visual manifestation of a particular phone is affected by its preceding

and following context. In order to model lip dynamics, we propose a *learning* framework in order to learn the parameters of a dynamic speech production model. We first record a training corpus of a human speaker uttering various sentences naturally, and obtain a low-dimensional parameterization of the lip shape using statistical shape-appearance techniques of {Jones98, Cootes98}. Motivated by the recent work of {Bridle99}, we model each phone as a target in lip space. Each target also comes with a so-called ``pliancy'', which determines how important it is to achieve that target during normal speech. Given a set of predetermined phone targets and pliancies for each, the actual observed lip parameters are instantiated using a dynamic forward-backward Kalman filter, which smoothly interpolates between the chosen targets, and efficiently takes into account the coarticulation effects. Finally we propose a learning algorithm to estimate the targets and pliancies for each phone that best explain the data.

2) Extend our system to use 3D models of faces and produce as output a 3D face complete of texture. The work will be done in collaboration with Thomas Vetter: we plan to explore the extension of our approach to the use of full 3D face models, following the approach of Vetter {VetterBlanz98}, itself an extension of Jones, Vetter and Poggio {VetterJonesPoggio97, {JonesPoggio96}. We plan to record a 3-dimensional face as it dynamically utters the same visual corpus we have designed, extract the visemes, and then morph between them to generate a synthetic 3dimensional talking head.

3) Can we synthesize a realistic video of a speaking person from just one image of the person? Because of recent results in our group and by our collaborators (Jones, Vetter...) we believe that this is possible. The question is an empirical one about the realism.

4) Extend our system to japanese. This would be a good collaborative project with a collaborator from NTT.

5) Incorporate higher-level communication mechanisms into our talking facial model, such as various expressions (eyebrow raises, head movements, and eye blinks).

6) Assess the realism of the talking face. We plan to perform several psychophysical tests to evaluate the realism of our system.

External Collaborators

Thomas Vetter of the Max Planck Institut in Tuebingen has worked for several years with our group at CBCL in the domain of graphics and more recently in

the specific domain of face synthesis. We plan to have Volker Blanz as a postdoc in this project starting after his PhD around fall 99.