Adaptive Man-Machine Interfaces
MIT9904-15

Progress Report: January 1, 2002—June 30, 2002

Tomaso Poggio

Project Overview

In this project we aim to achieve two significant extensions of our recent work on developing a text-to-visual-speech (TTVS) system (Ezzat, Geiger, Poggio 2002). The 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 our morphing approach from video to audio to **address issues of audio synthesis,** and 2) to extend the system to use morphing of **3D models** of faces -- rather than face images -- to output a 3D model of a speaking face.

The main applications of this work are for virtual actors, video dubbing, and 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.

An overview of our system is shown below:

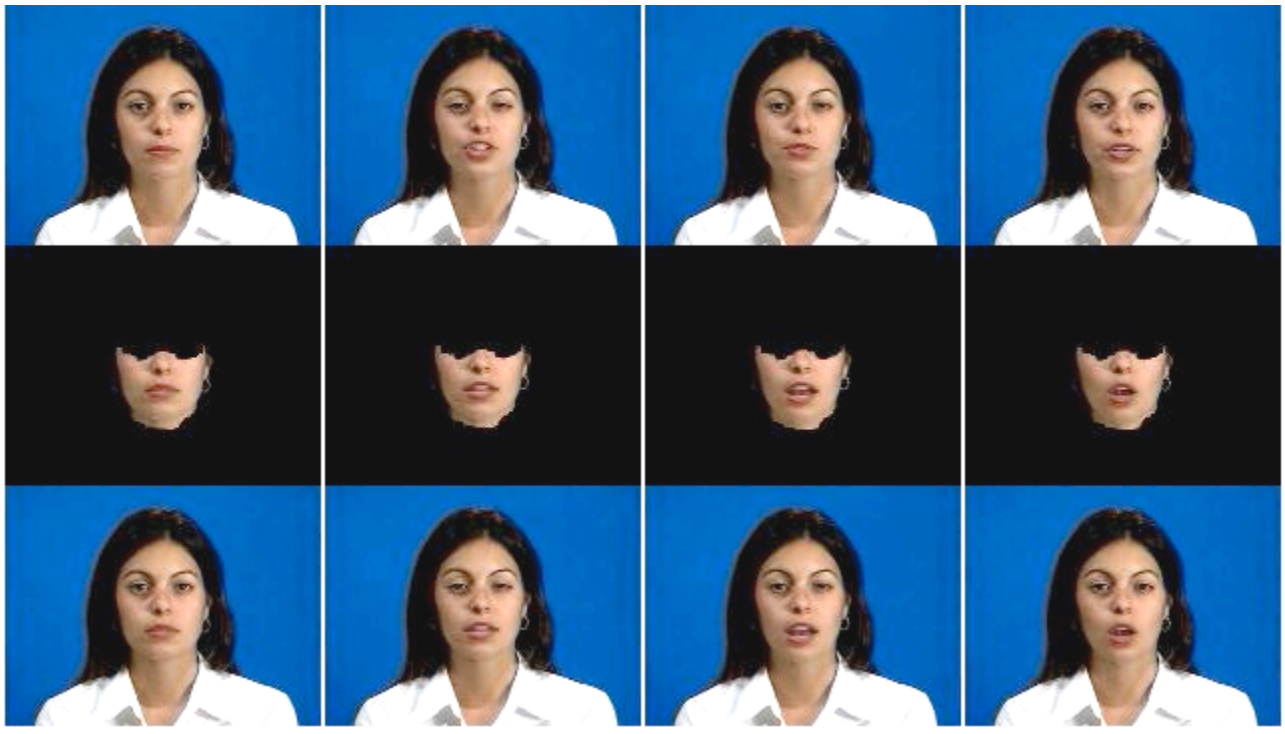


Figure 1: Top row: an original background sequence. Middle row: the synthetic mouth animations our system generates. Bottom row: the synthetic facial animations composited into the original background sequence.

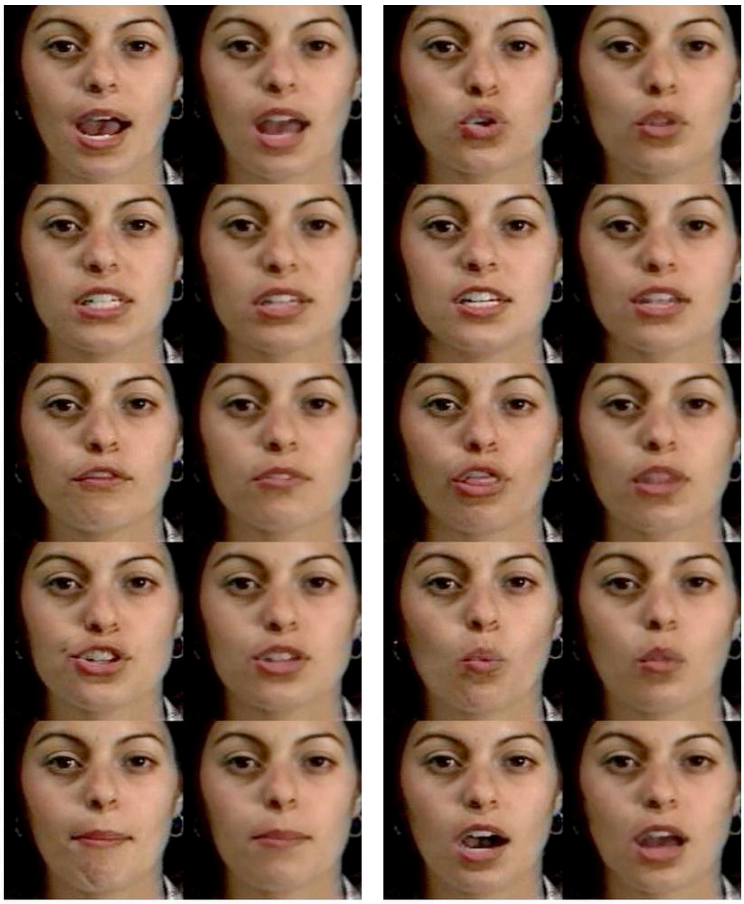
Progress Through June 2002

In the last six months we have completed our primary goal – a system for trainable videorealistic animation from images. **We have recently reported on this work in Ezzat, Geiger, Poggio 2002**.

In that work, we described how to create with learning techniques a generative, videorealistic, facial animation module. A human subject is first recorded using a digital videocamera as he/she utters a pre-determined speech corpus. After processing the corpus automatically, a visual speech module is learned from the data that is capable of synthesizing a visual stream of the human subject uttering entirely novel utterances that were not recorded in the original video. The output is videorealistic in the sense that it looks like a video camera recording of the subject. At run time, the input to the system can be either real audio sequences or synthetic audio produced by a text-to-speech system, as long as they have been phonetically aligned.

The two key contributions of this work were 1) an extension of the *multidimensional morphable model* (MMM) to synthesize new, previously unseen mouth configurations from a small set of mouth image prototypes; and 2) a *trajectory synthesis technique* based on regularization, which is automatically trained from the recorded video corpus, and which is capable of synthesizing trajectories in MMM space corresponding to any desired utterance.

Shown below are some of the synthetic facial configurations output by our system, along with their real counterparts for comparison:



Real Synthetic Real Synthetic

Figure 2 Some of the synthetic facial configurations output by our system, along with their real counterparts for comparison.

We also performed psychophysical tests to evaluate our synthetic animations. In particular, we performed two sets of "Turing tests" designed to test whether 22 naïve subjects could identify the synthetic animations from the real ones. In one experiment ("single presentation"), subjects were shown one animation, and asked to identify whether it was real or not. In the second experiment, subjects were shown two utterances, one real and one synthetic (but in randomized order), and asked to identify which one was real and which was synthetic. In both cases, performance was close to chance level (50%), and not significantly different from it.

Experiment	% correct	P<
Single presentation	54.3%	0.3
Double presentation	46.6%	0.5

Figure 3: Results from the visual "Turing tests"

Research Plan for the Next Six Months

We plan in the next six months to:

1) Begin extending our morphing approach from video to audio to **address issues of audio synthesis,**
2) Begin extending the system described above by acquiring **3D models** of faces -- rather than face images – in order to build a 3D morphable model capable of outputting 3D model of a speaking face.
3) Assess the **intelligibility** of the talking face by performing psychophysical tests similar to the visual "Turing test" ones.

References:

[1] Beymer, D. and Poggio, T. Image Representation for Visual Learning. Science, 272, 1905-1909, 1996
[2] Blanz, V., Vetter, T. A Morphable Model for the Synthesis of 3D Faces. In: Computer Graphics Proceedings SIGGRAPH'99, pp. 187--194, Los Angeles, 1999
[3] Ezzat, T. and T. Poggio. Visual Speech Synthesis by Morphing Visemes, International Journal of Computer Vision, 38, 1, 45-57, 2000.
[4] Ezzat, T. and T. Poggio. MikeTalk: A Talking Facial Display Based on Morphing Visemes. In: Proceedings of the Computer Animation Conference, Philadelphia, PA, 96-102, June 1998.
[5] Ezzat, T., Geiger G, and T. Poggio. Trainable Videorealistic Facial Animation. In Proceedings of SIGGRAPH 2002
[6] Jones, M. and Poggio, T. Multidimensional Morphable Models: A Framework for Representing and Matching Object Classes. Proceedings of the Sixth International Conference on Computer Vision, Bombay, India, 683 – 688, January 4-7, 1988
[7] Vetter, T. and Blanz, V. Estimating coloured 3d face models from single images: An example based approach. In Burkhardt and Neumann, editors, Computer Vision -- ECCV'98 Vol. II, Freiburg, Germany, 1998. Springer, Lecture Notes in Computer
Science 1407.