Adaptive Man-Machine Interfaces
MIT9904-15

Progress Report: Juy 1, 2001— December 31, 2001

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, 1998). 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 proposed 1) to 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) to address issues of coarticulation and dynamics. The main applications of this work are for virtual actors 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.

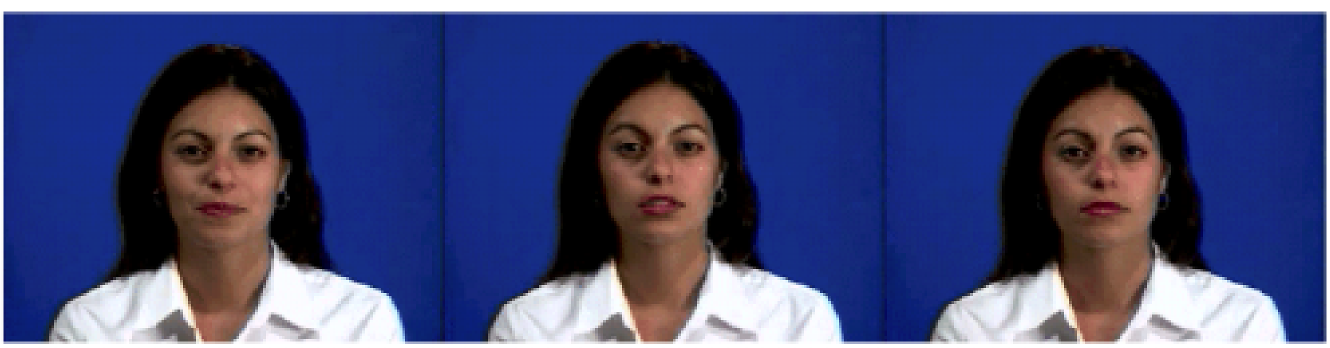


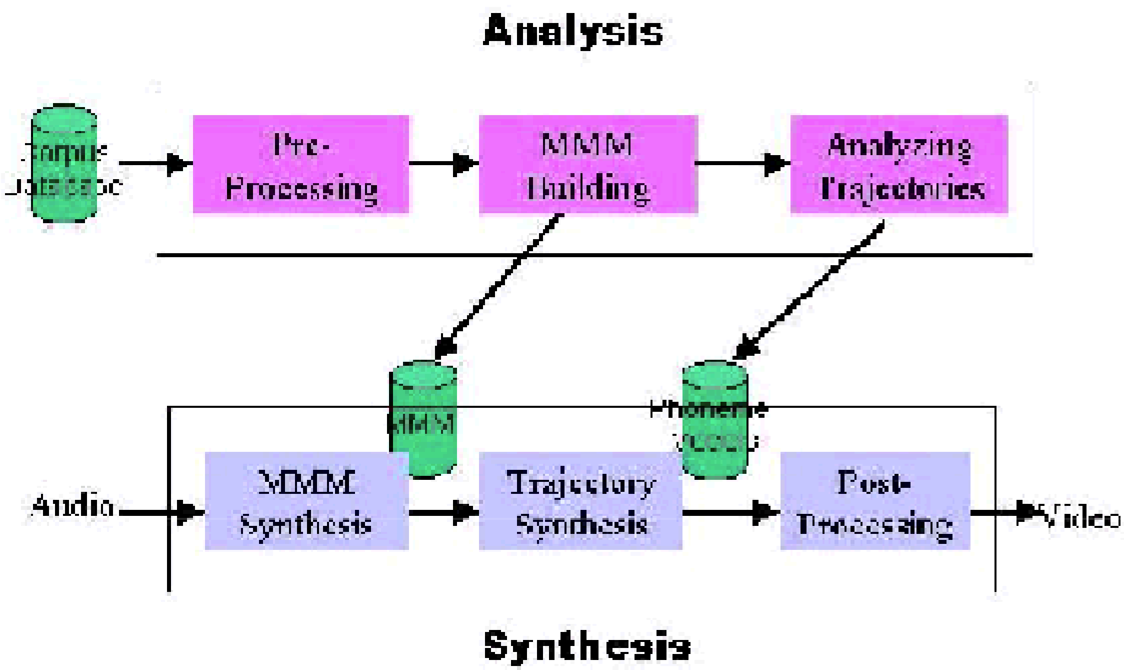
Figure 1

Progress Through December, 2001

In the last six months we have focused on the second goal – a system for trainable videorealistic animation from images. We have reported on the previous goal in the last report.

We describe how to create with learning techniques a genera-tive, 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 automati-cally, a visual speech module is learned from the data that is capable of synthesizing a visual stream of the human subject uttering en-tirely 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 sys-tem 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 paper are 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.



.

Figure 2: overview of the animation system

Figure 3: 24 of the 46 prototypes used to build the morphable model

We synthesize the trajectory by minimizing a regularization functional E:

$$E = \underbrace{(y-\mu)^T D^T (\Sigma^T \Sigma)^{-1} D (y-\mu)}_{target\_term} + \lambda \underbrace{y^T W^T W y}_{smoothness}$$

Coarticulation effects in our system are regulated via the magnitude of the variance for each phoneme. Small variance means the trajectory *must* pass through that region in phoneme space, and hence neighboring phonemes have little coarticulatory effect. Large variance means the trajectory has a lot of flexibility in choosing a path through a phonetic region, and hence it will choose to pass through regions which are closer to a phoneme's neighbors. The phoneme will thus experience large coarticulatory effects.

Research Plan for the Next Six Months

We plan in the next six months to:

1) Incorporate higher-level communication mechanisms into our (2D and possibly 3D) talking facial model, such as various expressions (eyebrow raises, head movements, and eye blinks).

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

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. and T. Poggio. Facial Analysis and Synthesis Using Image-based Models. In: Proceedings of the Workshop on the Algorithmic Foundations of Robotics, Toulouse, France, 449-467, August 1996.

[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.

[8] Vetter, T. and Poggio, T. Linear object classes and image synthesis from a single example
image. IEEE Transactions on Pattern Analysis and Machine Intelligence, 19(7):733--742, 1997.

[9] Vetter, T., Jones M.J. and Poggio, T. A bootstrapping algorithm for learning linear models of object classes. In: IEEE Conference on Computer Vision and Pattern Recognition – CVPR'97, Puerto Rico, USA, 1997. IEEE Computer Society Press.