Adaptive Man-Machine Interfaces
MIT9904-15

Progress Report: January 1, 2001— June 30, 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


Figure 2


Figure 3

Progress Through June 30, 2001

a) Mary101 – trainable videorealistic realistic animation from images

As part of our effort to develop a photo realistic talking face and learn facial dynamics from video, we have recorded a very large video corpus of a female subject. Images of the subject are shown in figures 1-3 above.

The corpus contains a large amount coverage of phonemes of the subject in 3 different emotional states: happy, neutral, and sad.

As the first of our efforts to learn facial dynamics, we have created a morphable model of the captured imagery. This morphable model, similar to ones described in [6][8], consists of a motion component and a texture component. Principal components analysis is performed on both the motion and texture components. This step allows us to represent the captured facial motion using a low-dimensional trajectories in the mouth motion space. Shown in Figure 4 below is one such trajectory for the sentence “More news in a moment” extracted by our morpheable model.

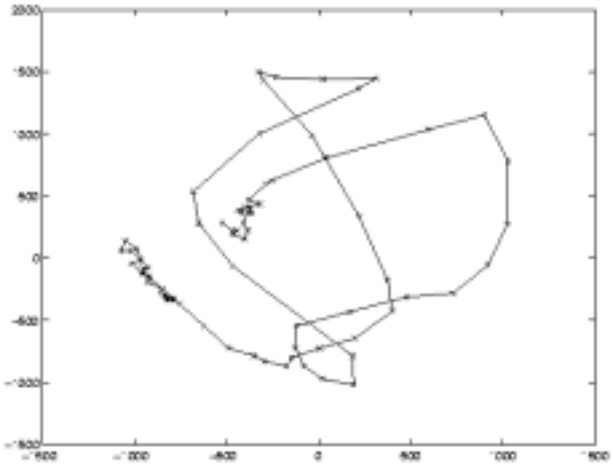

Figure 4: Trajectory for a sentence in our corpus
Extracted by our morpheable model.

Additionally, we have implemented a Hidden Markov Model framework in order to learn the dynamics and coarticulation smoothing effects of mouth motion. We have trained 40-50 phonemic HMMs on the data in our corpus. Each HMM is a 3-state left-to-right HMM, with Gaussian emission probabilities and diagonal covariances. Shown in Figures 5 and 6 below are the results of HMM training on two phonemes, /AY/ and /M/. The crosses represent the positional mean and covariance of the 3 states within each HMM. The pink trajectories represent actual trajectory data for each phone from our corpus. We also train 40-50 phonemic HMMs on velocity data as well.

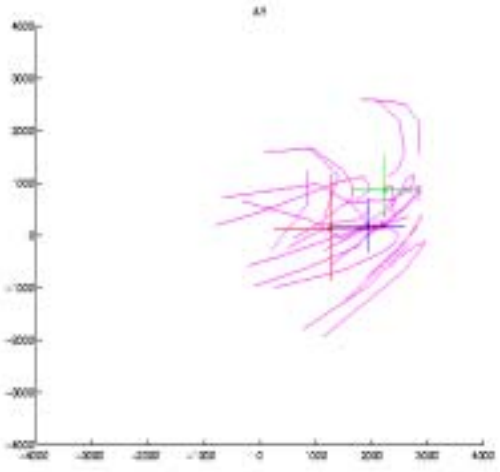

Figure 5

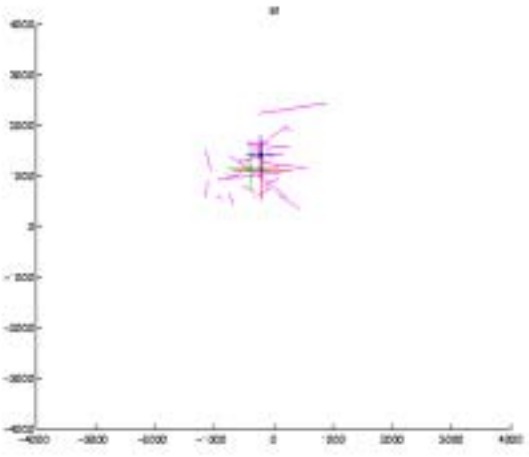

Figure 6

Using the trained HMMs we have begun to explore synthesis algorithms that generate trajectories from specified phonetic transcriptions. The current formulation of our synthesis algorithm concatenates the appropriate phonetic HMMs together, and samples from the HMMs in a smooth manner. The sampling is formulated as a curve generation algorithm that simultaneously tries to pass by the target points specified by the HMM output means while at the same time maintaining velocity and acceleration smoothness. Mathematically, our animation curve $x(t)$ is the solution of:

$$x(t) = \arg\min \sum_t (x-\mu)^T \Sigma^{-1} (x-\mu) + \sum_t (\dot{x}-\dot{\mu})^T \dot{\Sigma}^{-1} (\dot{x}-\dot{\mu})$$

where $\mu, \Sigma, \dot{\mu}, \dot{\Sigma}$ are the HMM output emission position and velocity means and covariances, respectively. Preliminary results from this algorithm currently synthesize underarticulated motion, however, and we are currently exploring how to remedy this problem.

3D Talking Face

In the previous report, we described a method to transfer the visual text-to-speech system developed by Ezzat and Poggio [3] to a three-dimensional face model. In order to make these results more widely applicable, we have collected new face data, developed a more flexible representation of face and teeth, made the data processing algorithms more precise and reliable, and finally transferred mouth movements from one individual to any other person. We recorded a new, considerably larger data set of 3D scans of a female face. The new set contains more of the relevant motions, more intermediate steps to help establishing correspondence, and a set of visemes (Figure 1 and Figure 2) which can be directly used to learn the motions that occur during speech. Due to the scanning hardware that we used, the scans were static, so the person had to sit still for about 10 seconds. Therefore, the laser scans of visemes might not be identical to what we see during natural speech. However, our data seem to reflect the natural movements of faces sufficiently well.

The process of establishing dense correspondence from all individual scans to a reference scan (Figure 3) has been further improved. Specifically, manual interaction has been reduced to only three simple tasks: (1) selection of a reference face, (2) manual marking of the teeth and the vertices in the inner part of the mouth on the reference scan, and (3) labelling of faces according to the degree of openness of the mouth. Based on this labelling, we applied a method similar to the bootstrapping approach described in [9] to add more and more variation to the model. Unlike [9], model dimensionality was not reduced by Principal Component Analysis during bootstrapping. Instead, starting from a reference face and a small set of relatively similar shapes, the system gradually included scans with more open and more closed mouths and thus covered more and more dissimilar shapes. We modified the representation of teeth and the dark, inner part of the mouth to make the system more flexible and to obtain more realistic results along the edges of the lips and teeth. As described in the previous report, we computed dense correspondence between facial surfaces, converted all scans to a vector space representation, and performed a Principal Component Analysis to find the main modes of variation in the data. We implemented a modeller program (Figure 4) to interactively combine and interpolate between different mouth poses, and finally generated animated video sequences.

With the new version of the system, it is now possible to transfer mouth movements from the prototype face to any other face (Figure 5 and Figure 6). We demonstrate this feature with an animated sequence of the person who was recorded for the system Mary101, and with the actor Tom Hanks. As described in [2], we recovered the 3D face shapes from single images, which was a frame from the video data for Mary101, and a frame from the motion picture ``Forrest Gump'' for Tom Hanks. Without any manual interaction, the scanned teeth and all the 3D displacements that occur during speech were transferred to the novel faces. It turned out that no rescaling of teeth or displacements was necessary for the particular faces that we worked with. Most likely, such a rescaling will be necessary for more dissimilar faces. However, this would be straight forward to implement, since the face space representation automatically provides 3D coordinates of fiducial points such as the corners of the lips or some points on the cheeks which can be used to obtain adequate scaling factors.

The results presented in this report demonstrate how motions of a person's mouth can be learned from a set of examples, and transferred to other individuals.

Figure 1: Opening and closing the mouth (bottom to top row) and changing the width of the mouth (left to right) are the most important degrees of freedom in mouth shape. The first part of our dataset of scans covers these variations.

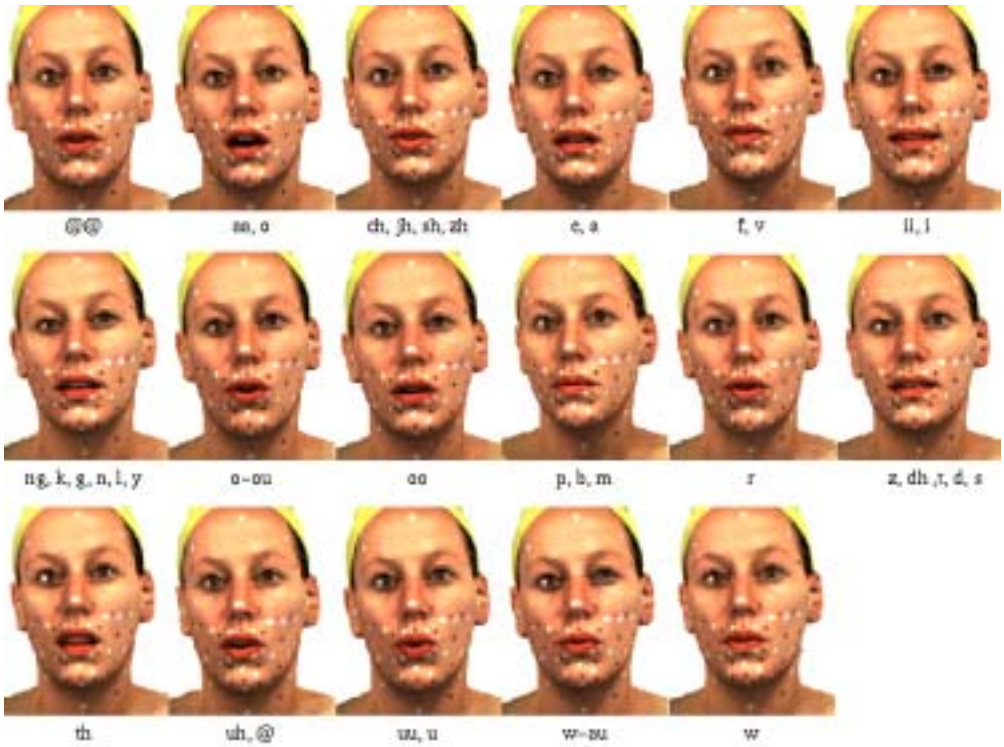

Figure 2 The second part of the database of 3D scans is formed by a set of visemes. These shapes are used to learn the specific changes in mouth shape during speech.

Figure 3: Left: This scan is used as a reference face: All other scans are matched with this intermediate mouth shape. Center: The teeth of the second scan are transferred to all other face shapes. The teeth's positions are fixed, relative to the upper half of the head. Right: The texture of this scan is used for generating motion sequences without the black and white markers drawn on the skin. After establishing correspondence, textures can be simply replaced.

Results:

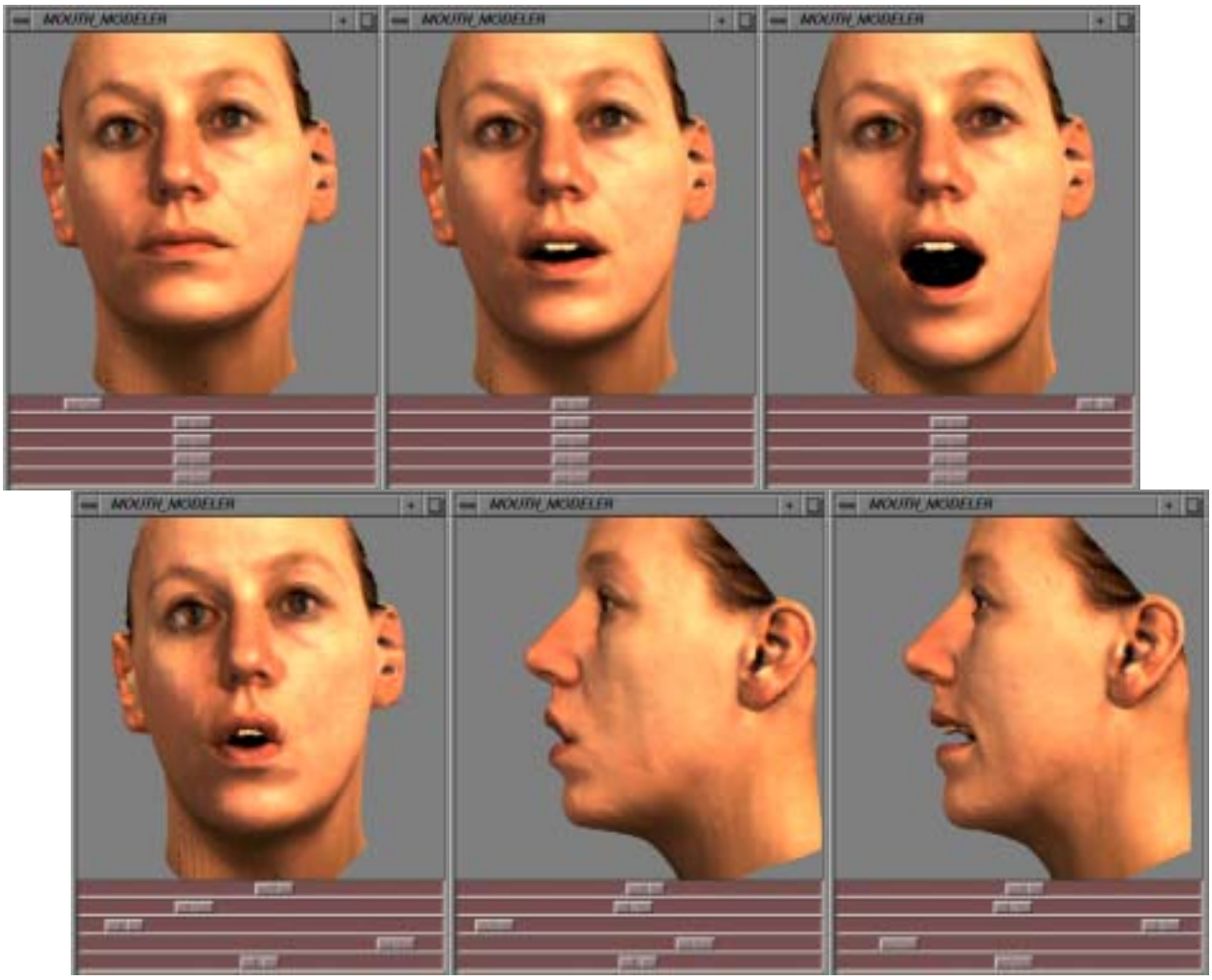

Figure 4 In an interactive modeler program, the user can control mouth shape with a set of sliders. The sliders are based on the orthogonal dimensions derived from a statistical analysis (PCA) of the database.

Figure 5 Given a single image of a person (left), we recover the 3D shape of the surface (second), and apply the vector displacements and color changes obtained from the initially scanned individual. Shapes and positions of teeth are transferred from the scanned individual (Figure 3). This modified face shape (third) can be composited into the original image (right).

Figure 6 Another example of face animation, based on a single image of an individual (left).

Research Plan for the Next Six Months

We plan in the next six months to:

1) develop further our HMM approach for dealing with the dynamics and coarticulation problem.
2) Explore a new approach for 3D talking faces using a realtime 3D scanner.
3) 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).
4) Assess the realism of the talking face. We plan to perform several psychophysical tests to evaluate the realism of our system.
5) Extend our approach using morphable model from TTVS to TTS. We plan to first study morphing of audio sequences. The system will take as input 2 audio sequences, and produce as output intermediate audio sequences that approximate natural exemplars lying between the 2 input sequences. Audio morphing might have important applications in speech synthesis.

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.