Adaptive Man-Machine Interfaces
MIT9904-15

Progress Report: June 30, 2000— December 31, 2000

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 **enrich the context** of each viseme to deal with coarticulation issues.

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.

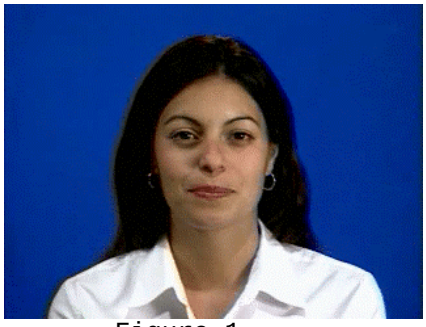


Figure 1

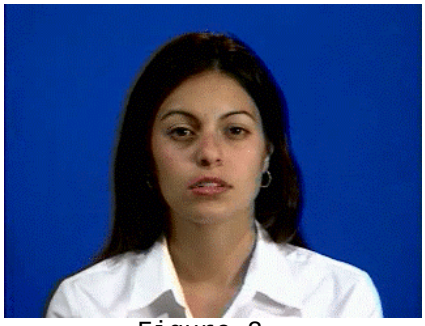


Figure 2

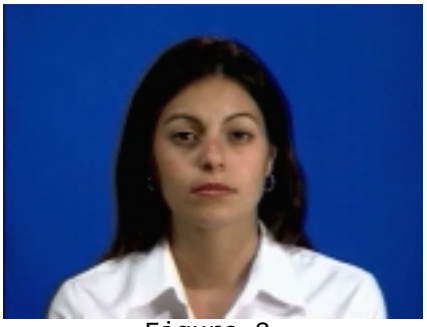


Figure 3

Progress Through December 2000

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.

Additionally, we have begun to explore a Hidden Markov Model framework in order to learn the dynamics and coarticulation smoothing effects of mouth motion. It is our belief that we can learn 40-50 phonemic HMMs and use them to resynthesize the observed mouth motions with the correct dynamics and coarticulation. Expected results using this approach will be demonstrated in March 2001.

b) 3D subproject

As planned, Volker Blanz and Tony Ezzat have worked to transfer the visual text-to-speech system developed by Ezzat and Poggio [3] to a three-dimensional face model. We do this by extending the approach proposed by Blanz and Vetter [2] to represent and manipulate three-dimensional face data.

The focus of the three-dimensional visual text-to-speech system presented here is to model deformations of the facial surface during speech in a realistic way. Muscle anatomy and the mechanical properties of tissue impose considerable constraints on the motions and poses that can be formed by human faces, and any violations of these natural constraints will be easily detected by viewers.

In our example-based approach ([1], [2], [3], [4], [5], [6]), we do not attempt to model the physical laws that control facial shape, but capture these shapes from a database of facial poses. Currently, our database consists of 19 three-dimensional laser scans of one face in different static poses. To simplify processing, black lines and dots are painted on the chin and cheeks. We use the scans to animate this particular face, and we can transfer the motions to novel faces in the future. It would be desirable to have three-dimensional scans of natural poses as they occur during speech, taken at short intervals while the person is speaking, or to have at least single instances from such a sequence. However, the scanning set-up that we currently employ does not provide such data, since recording a scan takes more than ten seconds. Still, the static poses in our database define all degrees of freedom of the mouth. To generate novel shapes during the animation, we form linear combinations of these scans, which are represented as face vectors in the framework of a Morphable Face Model ([2], [7], [8]).

In the project described here, we have implemented a method to transform the dataset into face vectors, and specifically, to establish dense correspondence between surface points on all individual scans. Dense correspondence is required whenever two three-dimensional shapes are morphed. In the Morphable Face Model, correspondence is used to consistently represent shapes and surface textures as face vectors, such that the x, y and z-coordinates of, say, the tip of the nose, are stored in the same three vector components in all face vectors. Thus, it is meaningful to form linear combinations of face vectors.

The face space of poses of the mouth is different in two respects from our previous work: First, the differences between a scan of an open and a closed mouth are challenging for the algorithm that establishes correspondence. Second, we have to provide a representation for the teeth and other structures that appear as the mouth is opened.

In order to ensure a natural appearance of the teeth, we do not attempt to design models of teeth and add them to the scans. Instead, we rely on the shape and colour information from one of the scans (m9) where the mouth is wide open. These teeth are used for all other poses, and whatever is visible from the teeth in other scans is discarded. The teeth remain fixed in space relative to the skull. For the upper jaw, this is anatomically valid. The lower jaw can move with respect to the skull, but as soon as the mouth closes slightly, the teeth are occluded by the lips, so this motion is hardly visible in the mouth region. In fact, movements of the lower jaw are more visible at the chin, a region that is modelled correctly by our approach.

Selected as our new reference face, the open mouth scan (m9) defines the topology of the polygonal mesh which will later be deformed to the shape of each individual scan, and to novel shapes. The 3D coordinates of the vertices of this mesh form the shape vector of the reference face. To compute all other shape vectors, we need to register the scans with the reference face.

On the face vectors derived from our dataset, we performed a Principal Component Analysis (PCA) to find the orthogonal vector dimensions with the largest variation within the dataset. We implemented a modeller program to from any direction. The sliders add or subtract principal components, and morph towards any of the original mouth poses from the database. We applied this program to model a set of 16 visemes. By morphing between these visemes, as described for a two-dimensional face model in [3], we generated video sequences from input text.

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.

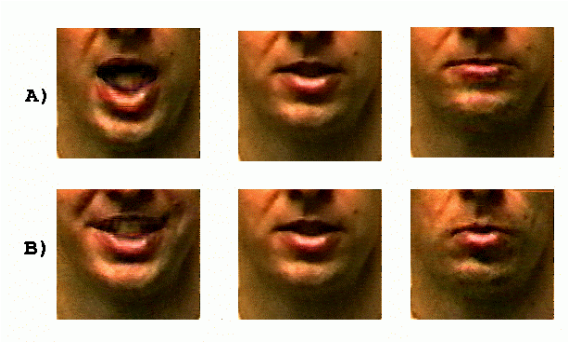
[8] Vetter, T. and Poggio, T. Linear object classes and image synthesis from a single example image. IEEE Transactions on Pattern Analysis and Machine Intelligence, 19(7):733--742, 1997.

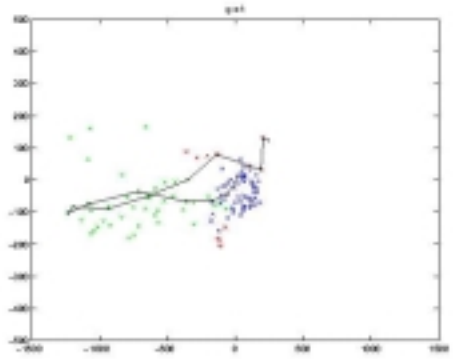
Figure 5: *A dataset of three-dimensional face scans defines the degrees of freedom of the Morphable Face Model. In 19 scans, black dots and lines are painted on the person's skin to disambiguate correspondence on cheek and chin. After correspondence was established, the textures of these scans were replaced by the texture of an additional scan (p3) that was recorded without make-up.*

Figure 6: A snapshot of the interactive modeller program. The face can be rotated, shifted and scaled. With the sliders, users can add or subtract principal components, or morph towards mouth poses from the database.

b) Coarticulation subproject

On the second subproject Tony Ezzat has made significant progress. He has recorded a training corpus of a human speaker uttering various sentences naturally, and obtained a low-dimensional parameterization of the lip shape using principal components (PCA). Shown in the image below on the left are the first two lip principal components: one axis represents degree of mouth opening and closing, while the other represents smiling-rounding. The technique also allows us to obtain trajectories of mouth movement, as shown in the image on the right. In general, we plan on using this technique to learn coarticulation models from video data, which may then be applied to 2D or 3D talking faces.





Research Plan for the Next Six Months

We plan in the next six months to:

1) develop further our approach to deal with the coarticulation problem. As we described we have recorded a training corpus of a human speaker uttering various sentences naturally, and obtained a low-dimensional parameterization of the lip shape. We will now use learning algorithms to generate the parameters of the morphable model from the phonetic time series, and which will implicitly incorporate coarticulation knowledge.

2) develop further the 3D talking face approach by recording more 3D visemes and morphing between them.