NTT-MIT Collaboration Proposal
Title: **Example-based image synthesis**.
PI: **Prof. William T. Freeman**
Date: Aug. 31, 2001

The capacities of disk drives and computer memories dramatically increase over time. Current capacities make feasible a variety of memory-intensive solutions to hard engineering problems. In this research, we will explore fast, memory-intensive algorithms for image synthesis and analysis, with application to realtime facial synthesis and super-resolution.

Real-time facial synthesis

Users of video cell phones may want to be able to choose the video persona that appears at the receiver's end. The displayed face could be that of a celebrity, or else the user's own face under favorable lighting, dress, and make-up. An option on cell phones could be the "output persona", and inclusion of this feature could provide a competitive advantage over other cell phone vendors or service providers.

However, the output display should reflect the expressions and motions of the user as he or she talks into the phone. To achieve that requires analyzing the speaker's face, and synthesizing a photo-realistic modified face in real time. Model-based approaches have been proposed [1], as well as some image-based rendering solutions [2], but photorealistic results have not been achieved in realtime.

A memory-intensive approach offers a solution to this problem. Long sequences can be recorded from a target output face. These sequences can be analyzed offline to find jump points that allow smooth transitions between sequences. This is in the spirit of the recent video textures work [3]. The sequences can also be translated or rotated into canonical positions or orientations to allow additional possible jump points.

We will drive the target output face from live video of the input face with a fast analysis of the motion and pose of the input face. At each instant, we will select which frame of the output face to display, striking a compromise between displaying natural motions of the output face (displaying them almost in the sequence that they appeared) and matching the input motions as well as possible (displaying the closest match to the input face).

As a specific example, using 2 or 3 hours of the video of former U.S. President Clinton's depositions, we should be able to synthesize photorealistic images of President Clinton's face, in his natural cadences, but reflecting the motions and expressions of a user seated in front of a camera.

Technical issues to be addressed in the research include: methods to prune and store the large dataset of the output faces. What method to analyze the input face will best maintain the expressions and motions of the input speaker in the re-rendered output? What is the best way to translate the style of the input person's facial motions into the style of those of the output person, which relates to our earlier work in analyzing style and content [4].

Super-resolution

The above application synthesizes output images at once from large regions of the training images (the whole face). Example-based image synthesis can also be applied to small regions of the image, as exemplified in a second application, super-resolution. A user wants to see a much higher resolution version of some received still or moving image. A conventional image interpolation algorithm, such as cubic spline interpolation, would give a very blurred image. This work would have application in image compression and enhancement, as well as for the enlargement of consumer digital or film photographs.

We have recently demonstrated a training-based method that gives much higher resolution image synthesis [5]. A large database stores many pairs of high and low resolution image patches. The input low-resolution image specifies a collection of high resolution candidate patches at each position. A compatibility constraint between neighboring high resolution patches is used to select the best candidate at each position. Researchers at the Kodak Imaging Science Technology Laboratory compared out algorithm with five other image interpolation algorithms, using eight naïve observers in preference test. Six out of the eight observers ranked our algorithm the highest.

However, in some images, artifacts occur, and to become practical this method needs to be artifact-free. We believe that the lack of local image information leads to erroneous synthesis outputs. We want to study combining a low-level estimate of the local region class with the local context information to select the best candidate high resolution patch. Furthermore, pruning of redundant entries in the database will allow more efficient use of training data. We hope these improvements make artifacts infrequent enough for practical use.

This super-resolution method has recently been generalized to allow fast texure synthesis [6, 7]. This opens the possibility for a variety of unexplored application areas using example-based image synthesis: combining object pose analysis with image synthesis to re-render a known set of objects under different lighting conditions, or in different appearance styles. These possible research directions could have applications to on-line shopping, or interactive games.

References:

[1] Darrell, T., Essa, I., and Pentland, A., "Task Specific Gesture Analysis using Interpolated Views", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, January, 1997.
[2] T. Ezzat and T. Poggio, "MikeTalk: A Talking Facial Display Based on Morphing Visemes", *Proceedings of the Computer Animation Conference* Philadelphia, PA, June 1998.
[3] A. Schodl, R. Szeliski, D. Salesin, I. Essa, "Video Textures", ACM SIGGRAPH, 2000.
[4] J. B. Tenenbaum, W. T. Freeman, "Separating style and content with bilinear models". *Neural Computation 12* (6), 1247-1283.
[5] W. T. Freeman, E. C. Pasztor, O. T. Carmichael, "Learning Low-Level Vision", *Intl. J. Computer Vision*, 40 (1), 25-47, 2000.
[6] A. Hertzmann, C. Jacobs, N. Oliver, B. Curless, and D. Salesin, "Image Analogies", ACM SIGGRAPH, 2001.
[7] A. A. Efros and W. T. Freeman, "Image Quilting for Texture Synthesis and Transfer", ACM SIGGRAPH, 2001.

TITLE: EXAMPLE-BASED IMAGE SYNTHESIS
PI: PROF. WILLIAM FREEMAN
SPONSOR: NTT
PROPOSAL BUDGET
9/1/01-6/30/03

			09/01/01 06/30/02	07/01/02 06/30/03	TOTAL
PERSONNEL	PERSON MONTH	EFFORT			
W. Freeman	1 MO/1.5 MOS	100.00 %	10,400	16,224	26,624
RA - PhD(2)	10 MOS/12 MOS	100.00 %	36,144	45,078	81,222
Total Salaries & Wages			46,544	61,302	107,846
Technical & Administrative Support 6.5%			3,864	4,857	8,721
Employee Benefits 18.0%			2,568	3,794	6,362
Vacation Accrual 9.5%			367	461	828
TOTAL PERSONNEL COSTS			53,343	70,414	123,757
OPERATING EXPENSES					
Travel			6,000	6,240	12,240
M & S			3,000	3,120	6,120
Communications			500	520	1,020
Publications			1,000	1,040	2,040
Network Charges			2,400	2,496	4,896
Network Facilities			3,600	3,744	7,344
Equipment			10,000	6,000	16,000
RA Tuition @ 35%			18,872	19,626	38,498
Allocated Expenses 3.8%			2,259	2,839	5,098
TOTAL OPERATING			47,631	45,625	93,256

EXPENSES

TOTAL DIRECT COSTS	**100,974**	**116,039**	**217,013**
OVERHEAD 65.5%	**40,162**	**50,853**	**91,015**
TOTAL CLOSING COSTS	**141,136**	**166,892**	**308,028**