Facial Expression Analysis and Synthesis

1. Research Team

Project Leader:	Prof. Ulrich Neumann, *IMSC and Computer Science*
Other Faculty:	Prof. Skip Rizzo, *IMSC and Gerontology*
Post Doc(s):	Reyes Enciso, John P. Lewis
Graduate Students:	Douglas Fidaleo, Junyong Noh
Undergraduate Students:	Albin Cheenath
Industrial Partner(s):	NCR

2. Statement of Project Goals

The Facial Expression project seeks to automatically record and analyze human facial expressions and synthesize corresponding facial animation. Analysis and synthesis of facial expression are central to the goal of responsive and empathetic human-computer interfaces. On the analysis side, the computer can respond and react to subtle sentiments reflected on the user's face. And on the synthesis end, the computer may present a comfortable and familiar visage to the end user. While the computer analysis of speech is the subject of extensive research, non-speech facial gestures have received less attention. Natural communication in virtual settings will require the development of a computational facility with facial gestures.

This work addresses both direct and indirect approaches to facial animation control. Direct methods frequently involve the transfer of 3D facial motion capture and are best suited to reproduction of realistic motion. Unfortunately, motion capture data is very tedious to acquire, and a costly calibrated multi-camera setup is required to estimate 3D positions of markers on the face. Our recent work on *3D Facial Motion Synthesis* involved learning of the facial motion subspace for estimation of 3D locations of 2D tracking data. This significantly lowers the cost and complexity for entry-level facial motion capture.

Realism, however, is often not the goal of animation. Indirect methods introduce a set of abstract parameters between the control signal and the animated character that allows model and animation independence. The animator can tailor the resulting animation with more artistic freedom. Facial gestures are a mid-level representation of facial state that span a representative portion of face space and can be abstracted into higher-level descriptions of facial state such as facial expressions.

3. Project Role in Support of IMSC Strategic Plan

Human-centric interfaces and interactions are part of IMSC's vision of Immersipresence and specifically important in the Communication Vision Project. These interfaces will require both the identification and processing of human facial gestures and (in the case of avatars) the synthesis of facial animation to match the user's speech. The Facial Expression project focuses on facial gesture identification and processing. A recent project, Data-Driven Facial Modeling and Animation, explores a data-driven approach to the synthesis problem. The two projects are complementary and the results will be merged into a single system.

4. Discussion of Methodology Used

Though practical to obtain, pixel intensity images of facial state are an unnecessarily high dimensional representation for a fairly low degree of freedom phenomenon (muscle contractions). To address this problem in the CoArt [3] work, the face is partitioned into local regions of change called *coarticulation regions* to constrain the number of muscle degrees of freedom in a given sample region. Unfortunately, it is very difficult for a human to label gesture data due to the close visual similarity of similar intensity gesture samples. Even with unsupervised clustering methods we cannot accurately order the clusters in terms of gesture intensity. Our early work was therefore limited to single gesture training data sets and simple template matching classification with features extracted by principal and independent component analysis.

By constraining the number of muscle degrees of freedom and reducing the dimensionality of gesture data with PCA, we uncover a coherent low dimensional structure to each gesture. This structure is modeled with second-order polynomials in *Gesture Polynomial Reduction (GPR)*. In addition to providing a concise and continuous model for expression space at the region level, the GPR representation enables extraction of continuous gesture intensity enabling fully parameterized performance driven facial animation.

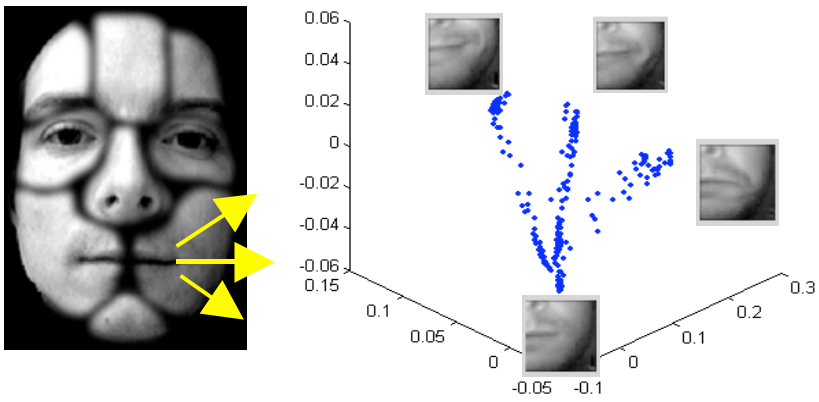

Figure 1. Gesture trajectories traced out by contraction of 3 independent muscles in the lower mouth region.

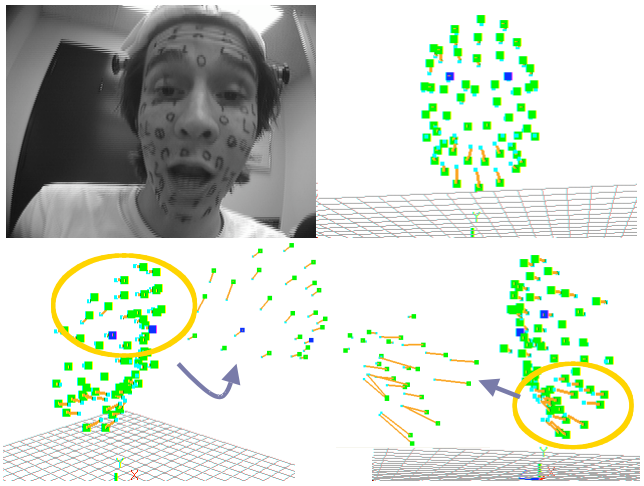

Figure 2. Face marked actor and synthesized 3D points viewed at different angles. Orange lines indicate 3D displacements.

Our new face-motion synthesis technique for direct animation control estimates 3D positions of markers with the convenience of using a single camera. The method first collects and analyzes possible trajectories of 3D face motions. Ideally, the data contains a variety of facial expressions and speech sequences. New 3D motions constrained in the learned face motion space can be synthesized from new 2D input. Initial training is performed utilizing data projection and clustering while the synthesis is done by 3D motion blending in response to new 2D inputs. The algorithm generates faithful 3D data for different environments from the training setting, actor/actress difference, marker placement variation, and camera parameter change. The estimated face motions look very natural, and a user study demonstrated that the motions are indistinguishable from actual 3D capture data.

5. Short Description of Achievements in Previous Years

An initial effort in 1999-2000 created a system that can rapidly build realistic face models suitable for use with gesture analysis [2]. Gesture analysis was used to classify the appearance of wrinkles and other dynamic features on the face. This was coupled with a 3D-texturing engine for reconstruction of wrinkle appearances on animated facial geometry. Subsequent improvements in the analysis work led to a more comprehensive partitioning of the face and an anatomic parameterization of the facial gestures (motivated by the set of facial muscles). As expressions can be easily defined in muscle space, gesture analysis at the muscle level allowed for interpretation of full facial expressions from input video.

Results from the analysis engine were connected to two animation modules. In the Emotion Driven Facial Animation module, abstract emotion parameters were analyzed from video sequences of an actor making facial gestures, and used to interpolate between hand generated 3D models. In the CoArt module, the lower-level muscle parameters were used to generate "flip-book" style 2D facial animation. Gesture analysis was applied to the mouth region independently to extract viseme information from the video and used to drive speech animation.

The Expression Cloning method was developed to transfer high quality facial animation data from one 3D model to another model with different geometry. Experiments were also performed applying motion frequency band decomposition to existing facial animation data for qualitative motion editing.

5a. Detail of Accomplishments During the Past Year

The GPR model for gesture activation in coarticulation regions was developed to overcome the limitations of the discrete face-state space defined in the CoArt system [3]. Extensive testing and validation of this model is being performed. The continuous input space deserves a corresponding continuous output space for visual validation. In response, the GPR has been connected to a novel 2D animation technique called Muscle Morphing where a mass-spring muscle system is used to drive a set of radial basis function control points.

Figure 3. (left and right) Expressions generated from contraction of mass spring muscles and corresponding deformation of a neutral image. (center) Neutral image with embedded musculature.

The 3D Face Motion Synthesis algorithm was completed and submitted to the SIGGRAPH 2003 graphics conference. Demonstrations include varied 3D data of expressions and speech sequences estimated from a set of synthetic and real 2D data. The validity of the 3D data was assessed by carefully designed quantitative and qualitative error measures. The results strongly indicate that the synthesized motions are indistinguishable from the captured 3D motions.

6. Other Relevant Work Being Conducted and How this Project is Different

Face tracking and recognition is a popular topic at present. The recognition of facial gestures, and in particular non-speech gestures including expressions of emotion, is not quite as mainstream but is being addressed by a number of groups. The bi-annual IEEE Conference on Automatic Face and Gesture Recognition is a good entry point to the current efforts in these areas.

Our work is distinguished in several ways. First, by dividing the analysis into regions, we can distinguish a large number of face states without entailing the combinatorial complexity of a whole-face classifier with the equivalent sensitivity. Second, the focus of our work includes non-speech gestures such as the “furled brow” (worried look as reflected in forehead wrinkles) as well as speech gestures. The mid-level analysis of the textural appearance of independent co-articulation regions into approximate underlying muscle activations appears to be a relatively unique approach at present. Existing gesture analysis work is focused on binary gesture classification and does not consider the intensity that is needed for animation or analysis of

expressions beyond the prototypical "universal" expressions defined by Paul Ekman.

Stereo cameras are often used to estimate an object position in 3D space by exploiting epipolar geometry. This general method is also applicable with markers on the face where multiple cameras, i.e. 6 ~ 8, are typically used to cover the entire face. Drawbacks of such a system include the expense, difficulty of camera calibration, synchronization, and missing and noisy data. Because of these difficulties, approaches using a single camera have been introduced as a viable alternative. However, the inherent limitation of 2D imagery often limits the range of 3D face motion space to predetermined morph targets, muscle configurations, or approximate 2D to 3D conversion using high level information such as MPEG4 facial animation parameters. In contrast, our method allows the faithful recovery of the relatively dense 3D face motion data so that professional animators can produce high quality animation.

7. Plan for the Next Year

The GPR representation has been explored only for person specific facial analysis. This is of primary importance for performance driven facial animation where we can readily acquire training data from the user. For ubiquitous computing applications, however, the analysis process must be generalized to unseen individuals. The GPR enables us to relate gesture training samples of arbitrary intensity across different subjects. This is a first step towards generalization of the gesture intensity classifier.

The appearance information analyzed in GPR is easy to acquire, can be processed efficiently, and encapsulates dynamic facial features such as skin wrinkling and bulging that is necessarily omitted by conventional point trackers and optical flow. However, due to significant differences in facial structure and transient facial features (eyebrows and wrinkles) those are difficult to normalize across subjects we hypothesize that strict appearance data is better suited for person specific analysis. Sparse motion information can be more easily normalized for geometric variations and will be explored to bootstrap the person-specific classifier for an unseen subject.

Current approaches to facial gesture classification, including our own work, are not 100% reliable. We will continue improving our techniques with the goal of increasing reliability and identifying a core set of effective algorithms. Goals also include the integration of face analysis with aural speech sensing in the Communication testbed. Neither domain has been entirely successful in understanding human speech, expression, or emotion. The merger of multiple modes is likely to produce improvements.

The estimated 3D data from the face motion synthesis algorithm has not yet been connected to animation. Most of the known motion capture animation techniques still demand a substantial amount of user intervention for high quality animation. Although Expression Cloning can serve to automate the process, the algorithm interpolates in a linear space and hence, degradation of animation quality can occur as the number of captured 3D data points decreases. We will investigate the use of higher order interpolation schemes especially around the mouth region where large deformation is required. Automatic detection of such an area and determination of the appropriate influence from an each 3D motion component will be a main focus.

Our current 3D synthesis algorithm is applied to the face domain. However, a similar framework should also be applicable for 3D body motion synthesis. A primary challenge in this domain will be dealing with self-occlusion while tracking the 2D points.

8. Expected Milestones and Deliverables

The Facial Gesture Analysis work is more exploratory than other IMSC projects directed towards human-centric computing. As such, the deliverables consist of a suite of programs developed for various facial tracking and gesture classification tasks, several experimental demonstrations of these programs at work, reports detailing the present state of the work [3-6] and the CoArt and Muscle Morphing 2D animation systems. A 10GB database of facial gesture data with expression intensity from ten individuals is also available.

The 3D motion synthesis deliverables include the system and a preliminary extendible database of motion training data. The Expression Cloning deliverables include the system, consisting of 36,000 lines of C++ program, and the reports [1, 2].

9. Member Company Benefits

With IMSC assistance, NCR has produced several internal reports on this technology and a video news piece that was disseminated to news broadcasters for inclusion in their programming. Demonstrations and video clips have been created for internal NCR use.

Non-verbal facial gestures are an important aspect of human communication. The ability to identify and reproduce non-verbal facial gestures will be needed in all forms of affective interfaces and virtual presence. The Facial Gesture Analysis project provides member companies with early access to both experimental investigations in this area and the tools needed to support such experiments.

Expression Cloning is a fundamental technique for avatars and for entertainment animation. Member companies working in these areas have access to this technique and its implementation at IMSC. Our current work on motion synthesis from 2D tracked feature points as well as gesture analysis provides a mechanism to generate high quality animation that can in turn be used by Expression Cloning. Member companies have access to the individual modules that can be assembled into a complete end-to-end system by which animation may be created by human performance and duplicated onto new facial models with minimal effort.

10. References

[1] J-Y. Noh, D. Fidaleo, U. Neumann, Gesture Driven Facial Animation, USC Technical Report 02-761, 2002.
[2] Junyong Noh and Ulrich Neumann, Expression Cloning, *ACM SIGGRAPH*, pages 277-288.
[3] D. Fidaleo and U. Neumann, Analysis of Coarticulation Regions for Performance Driven Facial Animation, *Journal of Visualization and Computer Animation*, 2003. (to appear)
[4] D. Fidaleo and U. Neumann, CoArt: Co-Articulation Region Analysis for Control of 2D/3D Characters, *Computer Animation* 2002.

[5] R. Enciso, J. Li, DA Fidaleo, T-Y. Kim, J-Y. Noh and U. Neumann, Synthesis of 3D Faces, Proceedings Digital and Computational Video 1999.
[6] D. Fidaleo, J-Y. Noh, T-Y Kim, R. Enciso, and U. Neumann, Classification and Volume Morphing for Performance-Driven Facial Animation, Proceedings Digital and Computational Video 1999.