M.I.T Media Laboratory Perceptual Computing Section Technical Report No. 370
Appears: Proceedings of Computer Animation '96 Conference, Geneva, Switzerland, June 1996

Modeling, Tracking and Interactive Animation of Faces and Heads using Input from Video

Irfan Essa, Sumit Basu, Trevor Darrell, Alex Pentland
Perceptual Computing Section, The Media Laboratory,
Massachusetts Institute of Technology
Cambridge MA 02139, U.S.A.
Email: {irfan|sbasu|trevor|sandy}@media.mit.edu

Abstract

We describe tools that use measurements from video for the extraction of facial modeling and animation parameters, head tracking, and real-time interactive facial animation. These tools share common goals but rely on varying details of physical and geometric modeling and in their input measurement system.

Accurate facial modeling involves fine details of geometry and muscle coarticulation. By coupling pixel-by-pixel measurements of surface motion to a physically-based face model and a muscle control model, we have been able to obtain detailed spatio-temporal records of both the displacement of each point on the facial surface and the muscle control required to produce the observed facial motion. We will discuss the importance of this visually extracted representation in terms or realistic facial motion synthesis. A similar method that uses an ellipsoidal model of the head coupled with detailed estimates of visual motion allows accurate tracking of head motion in 3-D.

Additionally, by coupling sparse, fast visual measurements with our physically-based model via an interpolation process, we have produced a real-time interactive facial animation/mimicking system.

Keywords: Facial Modeling, Facial Animation, Interactive Animation, Expressions and Gestures, Computer Vision.

1 Introduction

The communicative power of the face makes facial modeling and animation one of the most important topics in computer graphics. Originally, researchers focused on simply being able to accurately model facial motion [20, 27, 39, 22]. As the tools for facial modeling have improved, other researchers have begun to develop methods for producing extended facial animation sequences [17, 26, 41, 33]. The principle difficulty in both facial modeling and animation is the sheer complexity of human facial and head movement.

In facial modeling this complexity can be partially addressed by the use of sophisticated physical models of skin and muscle [33, 40, 26, 15]. However, there is very little detailed information on the spatial and temporal patterning of human facial muscles. This lack of information about muscle coactivation has forced computer graphics researchers to either fall back on qualitative models such as FACS (Facial Action Coding System, designed by psychologists [6] to describe and evaluate facial movements), or invent their own coactivation models [39, 29]. Consequently, today's best facial modeling employs very sophisticated geometric and physical models, but only primitive models of muscle control.

Lack of a good control model is also the limiting factor in production of extended facial animations. The best animations are still produced by artists who carefully craft key-frames [17, 8, 21], a time-consuming and laborious process. Even though the key-frame process does not require an explicit control model, it is likely that such a model would help by providing the artist with the right animation "control knobs."

The difficulty of facial animation has sparked the interest of the research community in *performance-driven* animation: driving a computer animation by simply animating your own face (or an actor's face). The VACTOR system [12], for instance, uses a physical system for measuring movement of the face. A system using infra-red cameras to track markers on a persons face has been reportedly used for several animations in movies. Williams [41] and Litwinowitcz [16] placed marks on peoples faces, so that they could track the 3-D displacement of the facial surface from video. Terzopoulos and Waters [34] used "snakes" [14, 32] to track makeup-highlighted facial features, and then used the displacements of these features to *passively* deform (*i.e.,* there was no interaction between the model and the measurements and the measurements just drive the model) a physically-based face model. Lee, Terzopoulos and Waters [15] have recently shown that the can generate very detailed 3-D facial models for animation. Saulnier *et al.* [30] suggest a template-based method for tracking and animation. Such methods have the limitations of requiring initial training or initialization, are limited in the range of face and head motions they can track, and are insensitive to very fine motions.

We feel that these automatic methods are an exciting direction in animation. However, current systems have several limitations. One limitation is their intrusiveness, as they require makeup, special cameras, or a physical probe. Another limitation is their relatively sparse spatial sampling, limiting the amount of detail that can be observed. A third limitation is that they model the face as a passive object, rather than as an actively controlled 3-D body.

On the other hand, little of work has been done on automatic extraction of head orientation from video. Extrac-

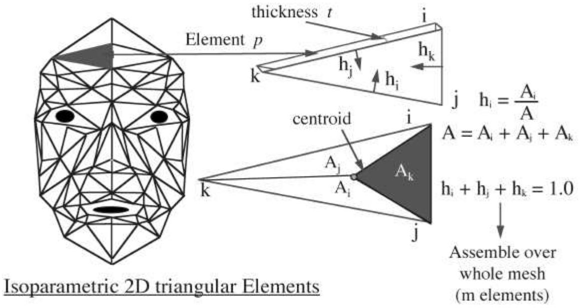


Figure 1: *Using the FACS mesh to determine the continuum mechanics parameters of the skin using FEM.*

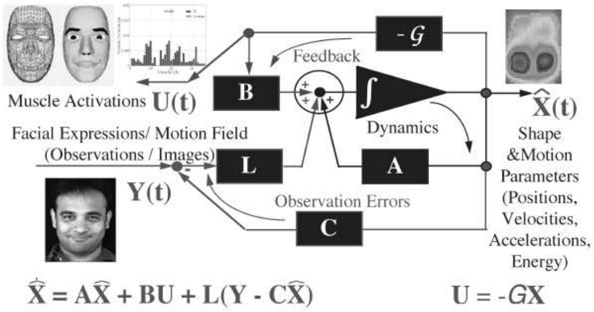


Figure 2: *Block diagram of the control-theoretic approach. Showing the estimation and correction loop (a), the dynamics loop (b), and the feedback loop (c) (from [11]).*

tion of head orientation is extremely important for human-machine interaction and for synthesis of a virtual actor with realistic head and facial motions. Azarbeyajani and Pentland [1] present a recursive estimation method based on tracking of small facial features like the corners of the eyes or mouth. However its use of feature tracking limited its applicability to sequences in which the same points were visible over the entire image sequence. Black and Yacoob [3] have developed a method that uses a eight parameter 2-D model for head tracking. Being inherently 2-D, this method does not allow estimation of 3-D parameters.

Our Approaches

In this paper we attempt to improve on these previous systems by removing the need for surface markings, allowing more detailed geometric measurement, and by formulating the problem in an active control framework for detailed analysis of facial motion. We will also describe two additional tools; one for tracking of heads from video and a another as a real-time extension for interactive facial animation. This will extend our detailed extraction of facial actions method by using coarse measurements from video to guide the graphics process. These tools share common goals but rely on varying details of physical and geometric modeling and in their input measurement system. We discuss them briefly here:

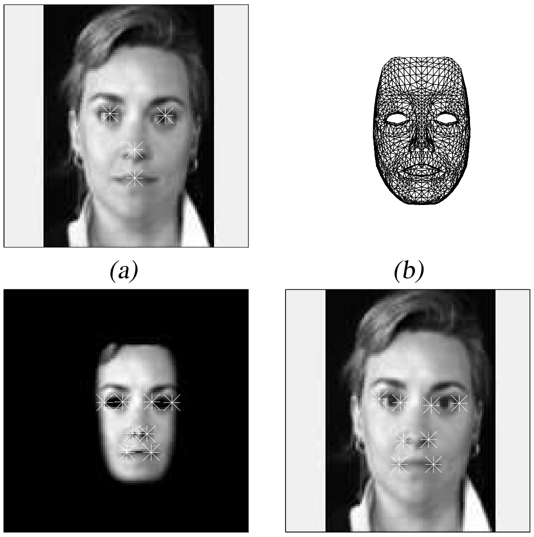
Facial Modeling and Analysis: Modeling facial motion requires detailed measurement across the entire facial surface. Consequently, our facial modeling tool uses pixel-by-pixel measurements of surface motion (*optical flow [13]*) as input measurements. These dense motion measurements are then coupled to a *physically-based face model* and to a *muscle control model*. The outputs of this modeling process are detailed records of both the displacement of each point on the facial surface, and the muscle control required to produce the observed facial motion. The recovered and muscle control patterns can be used to animate other models or composed to make new combination expressions.

The advantage of this approach over *a priori* facial modeling is that we can observe the complex muscle coarticulation patterns that are characteristic of real human expressions. For instance, it has been observed that a major difference between real smiles and forced or fake smiles is motion near the corner of the eye [5]. We have been able to observe and quantify the relative timing and amplitude of this near-eye motion using our system.

Interactive Animation: Ideally, we would like to use the above method for interactive animation. However, the above method extracts fine-grained information and is hence far from "interactive-time." We also describe a system for interactive facial animation that builds on our detailed extraction of facial patterns. Facial animation typically involves sequencing a relatively small set of predetermined facial expressions (e.g., lip smiling, pursing, stretching, and eye, eyelid, and eyebrow movement) rather than determining the motion of each facial point independently. That is, there are often relatively few independent geometric parameters each of which may have a large amount of temporal variation.

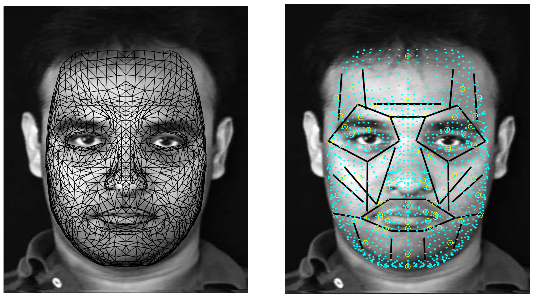
Consequently, we can use fairly simple visual measurements to establish the geometric parameters, but we would like to do this very quickly — in real time if at all possible. In our real-time system the visual measurements are normalized correlation between the face and a small set of pre-trained 2-D templates. This type of measurement has the advantage of both being very robust and fast; we use commercial image processing hardware (from Cognex, Inc.) so that the image measurement process can occur at frame rate. These measurements are then coupled with our physically-based model's parameters via an interpolation process, resulting in a real-time (passive, *i.e.,* the observations drive the model) facial animation system.

Head Tracking and Orientation: Since head orientation plays a major role in both analysis and synthesis of facial movements, we also introduce a method for robust tracking of head movements in extended video sequences. This method is based on regularization of optical flow using a 3-D head model for robust and accurate tracking in 3-D using only a single camera. This model-based method does not require the same features on the face to be visible over the entire length of the sequence and is stable over



(a) (b) (c) (d)

Figure 3: *Initialization on a face image using methods described by Pentland et al. [19, 24], using a canonical model of a face.*



(a) Mesh (b) Muscles

Figure 4: *(a) Face image with a FEM mesh placed accurately over it and (b) Face image with muscles (black lines), and nodes (dots).*

extended sequences, including those with large and rapid head motions. Additionally, this method allows tracking of all the six degrees of freedom of the rigid motion of the head, dealing gracefully with the motion singularities that most template-based methods fail to handle.

Due to space considerations we will not go into the details of the visual measurement techniques or the the physics-based modeling of the face. Referenced work provide adequate technical details.

2 Facial Action Parameters Extraction

For detailed estimation of facial action parameters, we need to develop a detailed physics-based model of a face. We use a polygonal model of a face to generate a finite element mesh (Figure 1). The interpolation and strain-

Figure 5: *Expressions from video sequences for various people in our database. These expressions are captured at 30 frames per second at NTSC resolution (and cropped appropriately. We have by now developed a video database of over 30 people under different lighting conditions and backgrounds. We are also incorporating head movements into our database.*

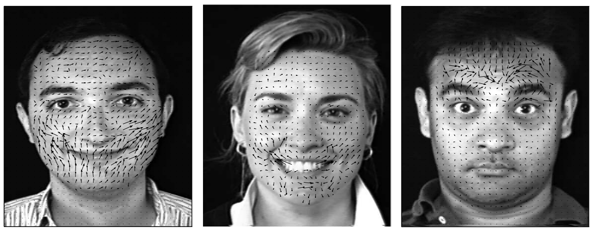


Figure 6: *Motion fields for few of the observed expressions.*

displacement matrices are determined for the given geometry by using the triangular polygons as two dimensional isoparametric shell elements. Using these interpolation and strain-displacement matrices, the mass, stiffness and damping matrix for each element were computed and then assembled into a matrix for the whole mesh [9].

Material properties of real human skin were used in this computation. These material properties were obtained from Pieper's Facial Surgery Simulator [26] and from other studies on mechanical properties of human bodies (*e.g.,* [37]). Physically-based skin model muscles were attached to this using the method of Waters and Terzopoulos [39, 33] and using muscle data from Pieper [26]. This provides an anatomical muscle model of the face that deforms on actuation of muscles. We believe that this is an extremely detailed model for facial animation. However, this model is unable to represent wrinkles as Viaud *et al.* [35] models.

Our facial modeling system functions by using optical motion measurements to drive the physical face model. However, such measurements are usually noisy, and such noise can produce a chaotic physical response. Consequently an estimation and control framework needs to be incorporated into the system to obtain stable and well-proportioned results [11].

To begin analysis of a facial motion, the geometric mesh needs to be initialized and accurately fit to the face in the

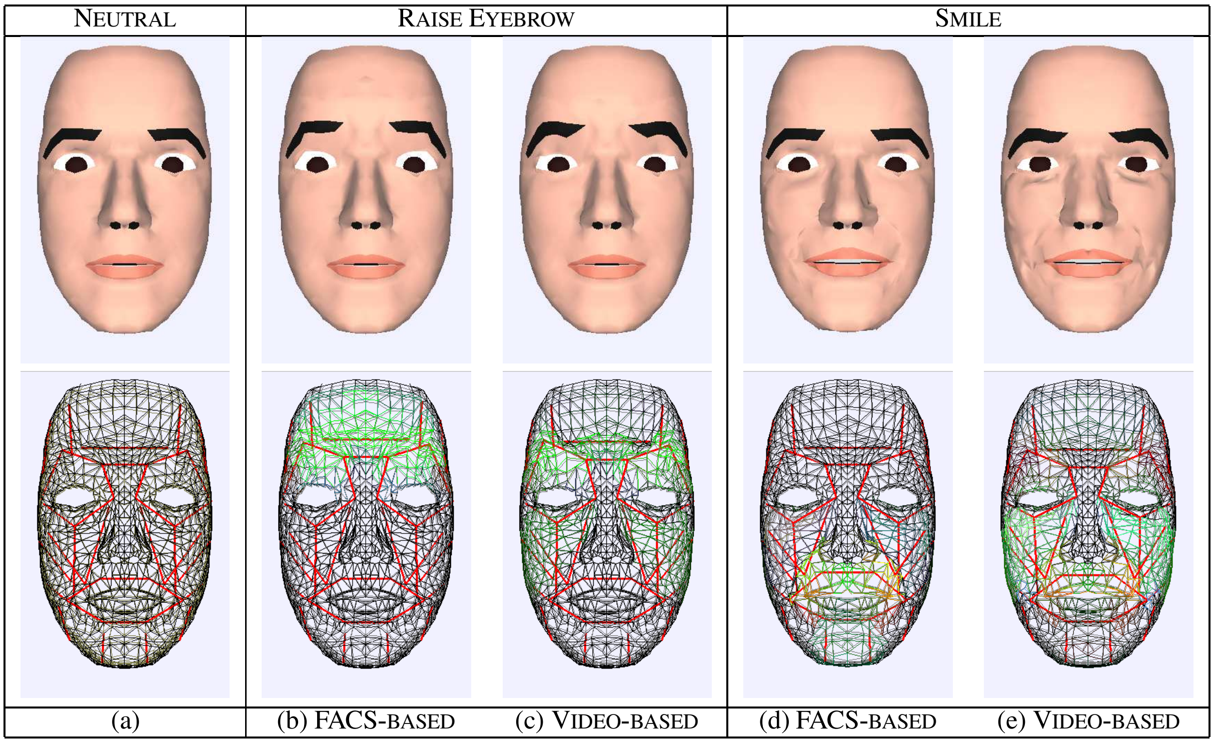


Figure 7: *Neutral Face (shaded/wireframe). Columns (b) and (d) animations produced by FACS model and Columns (c) and (e) visual measurements for raise eyebrow and smile expressions.*

image. For this we need to locate a face and the facial features in the image. To automate this process we are using the View-based and Modular Eigenspace methods of Pentland and Moghaddam [18, 25].

Using this method we can automatically extract the positions of the eyes, nose and lips in an image as shown in Figure 3 (a). These feature positions are used to warp the face image to match the canonical face mesh (Figure 3 (b) and (c)). This allows us to extract the additional "canonical feature points" on the image that correspond to the fixed (nonrigid) nodes on our face mesh (Figure 3 (d)).

After the initial registering of the model to the image as shown in Figure 4, Pixel-by-pixel motion estimates ("optical flow") are computed using methods of Simoncelli [31] and Wang [38]. The model on the face image tracks the motion of the head and the face correctly as long as there is not an excessive amount of head movement during an expression. Motion vectors for some of these expressions are shown in Figure 6. This motion is projected onto the mesh and produces deformation of the skin. The control-feedback loop (see Figure 2) estimates the muscle control needed to produce the observed temporal and spatial patterning. Mathematical details of the model and estimation framework are described in [9].

Limitations of Existing Representations

The goal of this work is to develop a tool for more accurately modeling facial motion. The current state-of-the-art for facial description (either FACS itself or muscle-control versions of FACS) have two major weaknesses:

- The action units are purely local spatial patterns. Real facial motion is rarely completely localized; Ekman himself has described some of these action units as an "unnatural" type of facial movement.

- There is no time component of the description, or only a heuristic one. From EMG studies it is known that most facial actions occur in three distinct phases: *application*, *release* and *relaxation*. In contrast, current systems typically use simple linear ramps to approximate the actuation profile.

Other limitations of FACS include the inability to describe fine eye and lip motions, and the inability to describe the coarticulation effects found most commonly in speech [7, 22]. Although the muscle-based models used in computer graphics have alleviated some of these problems [33], they are still too simple to accurately describe real facial motion.

Consequently, we have focused our efforts on characterizing the functional form of the actuation profile, and

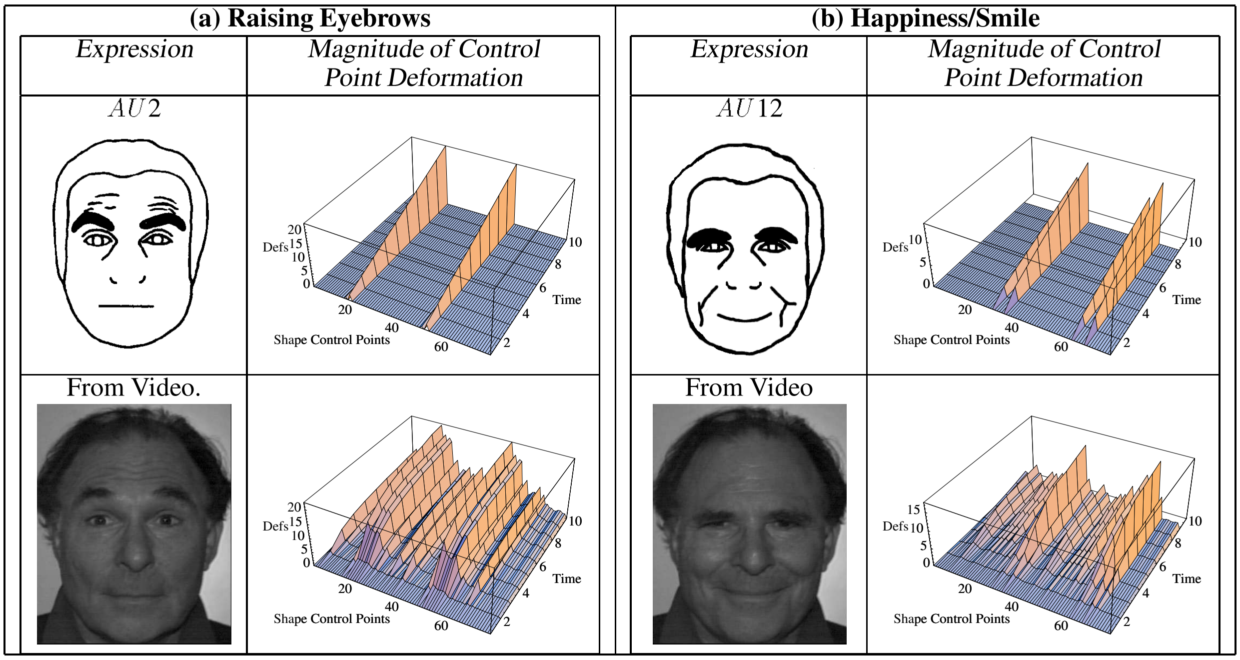


Figure 8: *FACS/CANDIDE deformation vs. Observed deformation for the (a) Raising Eyebrow and (b) Happiness expressions. Surface plots (top) show deformation over time for FACS action (a) $AU2$ and (b)$AU12$, and (bottom) for an actual video sequence of raising eyebrow and happiness.*

on determining a basis set of "action units" that better describes the spatial properties of real facial motion. We will illustrate our results using the smile and eyebrow raising expressions.

Facial Motion Measurements

The first step in modeling facial motion from video data is to acquire image sequences of subjects making expressions. For this purpose we arranged a video taping of over 30 subjects making expressions. All of the results that are described here are based on this video data. Some of the frames of these sequences are shown in Figure 5.

After digitizing the acquired video sequences Figure 5, optical flow were computed for the actions of raising eyebrow, lowering eyebrow, lip tightening and smile, frown, surprise, anger, disgust and a series of other expressions.

These dense motion measurements were then fed into the control feedback loop shown in Figure 2, and muscle activations produced. This step, when coupled with the motion estimation method, results in analysis of expressions at the rate of 60 seconds/frame on an SGI Onyx Reality Engine workstation, using only one processor. Now we will briefly discuss the validity of our analysis and modeling.

Resulting models of facial motion

The first column of Figure 7 shows the model in neutral (relaxed) state. The second shows the expressions as generated by using a standard FACS implementation and then using our representation extracted from video for the raise eyebrow expressions. The last column shows the generated expressions of smile using FACS and our representation. For our standard FACS implementation, we are using the CANDIDE model, which is a computer graphics model for implementing FACS motions [29].

To illustrate that the resulting parameters for facial expressions are more spatially detailed than FACS, comparisons of the expressions of *raising eyebrow* and *smile* produced by standard FACS-like muscle activations and our visually extracted muscle activations are shown in Figure 9. As expected, the two models are very similar in the primary FACS activation region. For the case of eyebrow raising, both models are similar in the area directly above the eyebrow. For the smiling example both models are similar in the area immediately adjacent to the mouth.

In both cases, however, the visual measurement model had significant additional deformations in distant areas of the face. In the case of eyebrow raising, the visual model has additional deformations high in the forehead, immediately above the eye, and in the lower cheek. In the case of smiling, there are additional deformations beneath and between the eyes, on the far cheek to either side of the mouth, and on the temples. These differences are explored in more detail in the following sections.

Spatial patterning: The top row of Figure 8 shows $AU2$ ("raising eyebrows") and $AU12$ from the FACS model and a linear actuation profile for the corresponding geometric control points. This is the type of spatial-temporal patterning commonly used in today's computer graphics anima-

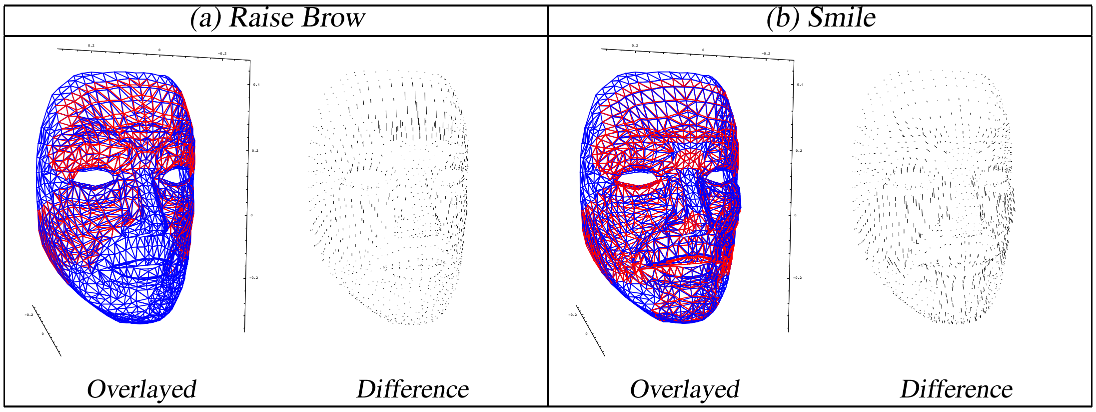


Figure 9: *Vision-based expression overlayed on top of a FACS expression of (a) Raising Eyebrow, and (b) Smile, and the differences between the two facial motions. Red shows excessive motion on the surface as modeled my visually extracted parameters. It can be seen that modeling by visual measurement produces a more detailed pattern of motion.*

tions. Below this is shown the observed motion of these control points for the expressions of raising eyebrows (labeled by Ekman as $AU2$) and smile (labeled by Ekman as mostly $AU12$). As can be seen, the observed pattern of deformation is very different than that assumed in the standard computer graphics implementation of FACS. There is a wide distribution of motion through all the the control points, and the temporal patterning of the deformation is far from linear. By using these observed patterns of motion, rather than the simple actuations typically assumed, more realistic computer animations can be produced.

Temporal Patterning: Figure 10 shows plots of facial muscle actuations for the smile and eyebrow raising expressions. In this figure the 36 muscles were combined into seven local groups for purposes of illustration. As can be seen, even the simplest expressions require multiple muscle actuations.

Of particular interest is the temporal patterning of the muscle actuations. We have fit exponential curves to the activation and release portions of the muscle actuation profile to suggest the type of rise and decay seen in EMG studies of muscles. From this data we suggest that the relaxation phase of muscle actuation is mostly due to passive stretching of the muscles by residual stress in the skin.

Note that Figure 10(b) also shows a second, delayed actuation of muscle group 7 about 3 frames after the peak of muscle group 1. This example illustrates that coarticulation effects can be observed by our system, and that they occur even in quite simple expressions.

3 Interactive Facial Animation

Because face models have a large number of degrees of freedom, facial modeling requires dense, detailed geometric measurements in both space and time. Currently such dense measurement is both computationally expensive and noisy; consequently it is more suitable to undertake off-line analysis of discrete facial movements rather than real-time

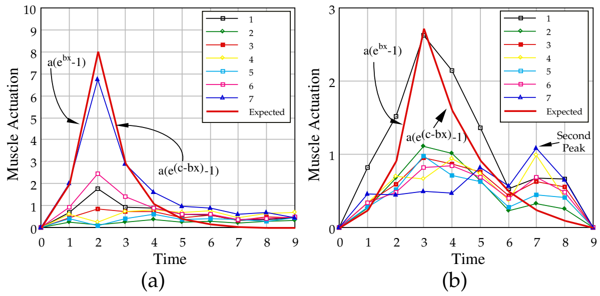


Figure 10: *Actuations over time of the seven main muscle groups for the expressions of (a) raising brow, and (b) smile. These plots shows actuations over time for the seven muscle groups and the expected profile of application, release and relax phases of muscle activation.*

analysis of extended facial action. Facial animation, in contrast, typically involves temporally sequencing between a fixed set of predefined facial actions. For instance, an animation sequence might consist of the lip movements associated with speech plus a few eye motions plus eyeblinks and eyebrow raises.

Because the full range of facial motion is typically not present in any particular animation sequence, the number of degrees of freedom required for the animation is limited. One can think of the animation as having a fixed, relatively small set of "control knobs," one for each type of motion, and then producing the animation by moving these control knobs appropriately. As described in the previous section, the muscle parameters associated with these control knobs are determined by the off-line modeling of each individual type of facial action.

The major question, of course, is when and how much to move each control knob (face control parameter). In our system the setting of each muscle control parameter is de-

Figure 11: *2-D Full-Face templates of neutral, smile and surprise expressions used for tracking facial expressions. See Figure 13 and Figure 14(a).*

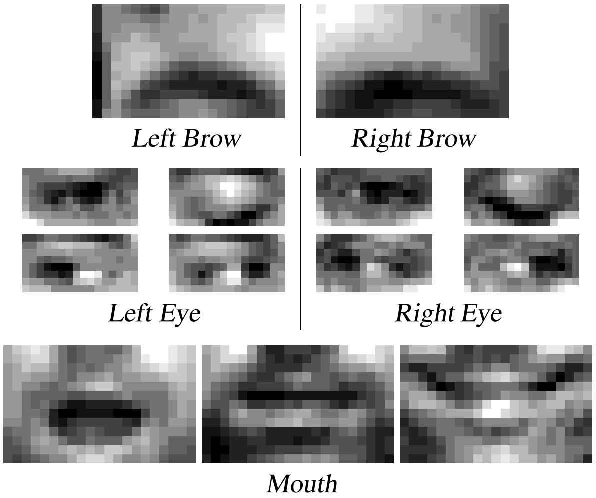


Figure 12: *2-D Eye-brows [Raised], Left and Right Eyes [Open, Closed, Looking Left, and Right], and Mouth templates [Open, Closed and Smiling] used for tracking facial expressions. These images are showing the exact resolution as used by the hardware. Blur your vision to see it the real details. See Figure 14(b).*

termined using sparse, real-time geometric measurements from video sequences.

One way to obtain these measurements would be to locate landmarks on the face, and then adjust the control parameters appropriately. The difficulty with this approach is first that landmarks are difficult to locate reliably and precisely, and second that there are no good landmarks on the cheek, forehead, or eyeball.

An alternative method is to teach the system how the person's face looks for a variety of control parameter settings, and then measure how similar the person's current appearance is to each of these known settings. From these similarity measurements we can interpolate the correct control parameter settings. In our experience this method of determining control parameters is much more robust and efficient than measuring landmark positions.

Our similarity metric is simply the correlation of a previously stored intensity view with the new data. We take views corresponding to each trained expression for which we have obtained detailed force and timing information using the method outlined in the previous section. By constraining the space of expressions to be recognized, we can match/recognize existing expressions rather then derive new force controls for the input video, and dramatically improve the speed of the system.

When the input image matches one of the trained examples, the corresponding previously stored motor controls are actuated in the facial model. If there is no match between the image and the existing expressions, an interpolated motor actuation is generated based on a weighted combination of expressions. The mapping from vision scores to motor controls is performed using piecewise linear interpolation implemented using a Radial Basis Function (RBF) network [28]. (We have also implemented a Gaussian RBF and obtained equivalent results.) This specific implementation with details on learning and interpolation techniques and the appearance-based method for both faces and hands is explored in much detail in [10, 4].

The RBF training process associates the set of view scores with the facial state, *e.g.,* the motor control parameters for the corresponding expression. If we train views using the entire face as a template, the appearance of the entire face helps determine the facial state. This provides for increased accuracy, but the generated control parameters are restricted to lie in the convex hull of the examples. View templates that correspond to parts of the face are often more robust and accurate than full-face templates, especially when several expressions are trained. This allows local changes in the face, if any, to have local effect in the interpolation.

Figure 11 shows the eye, brow, and mouth templates used in one of the examples in the videotape. The normalized correlation calculation is carried out by commercial image processing hardware from Cognex, Inc. The normalized correlation matching process allows the user to move freely side-to-side and up-and-down, and minimizes the effects of illumination changes. The matching is also insensitive to small changes in viewing distance ($\pm 15\%$) and small head rotations ($\pm 10^\circ$).

For each incoming frame of video, all of these 2-D templates are matched against the image, and the peak normalized correlation score recorded. Note that the matching process can be made more efficient by limiting the search area to near where we last saw the eye, mouth, etc.

Experiments: Figure 13 illustrates an example of real-time face animation using this system. Across the top, labeled (a), are five video images of a user making an expression. Each frame of video is then matched against all of the templates shown in Figure 11, and normalized correlation scores are measured. A plot of the normalized correlation score for each template is shown in (b). These scores are then converted to state estimates and fed into the muscle control loop, to produce the muscle control parameters shown in (c). Five images from the resulting animation sequence are shown in (d). Figure 14 shows

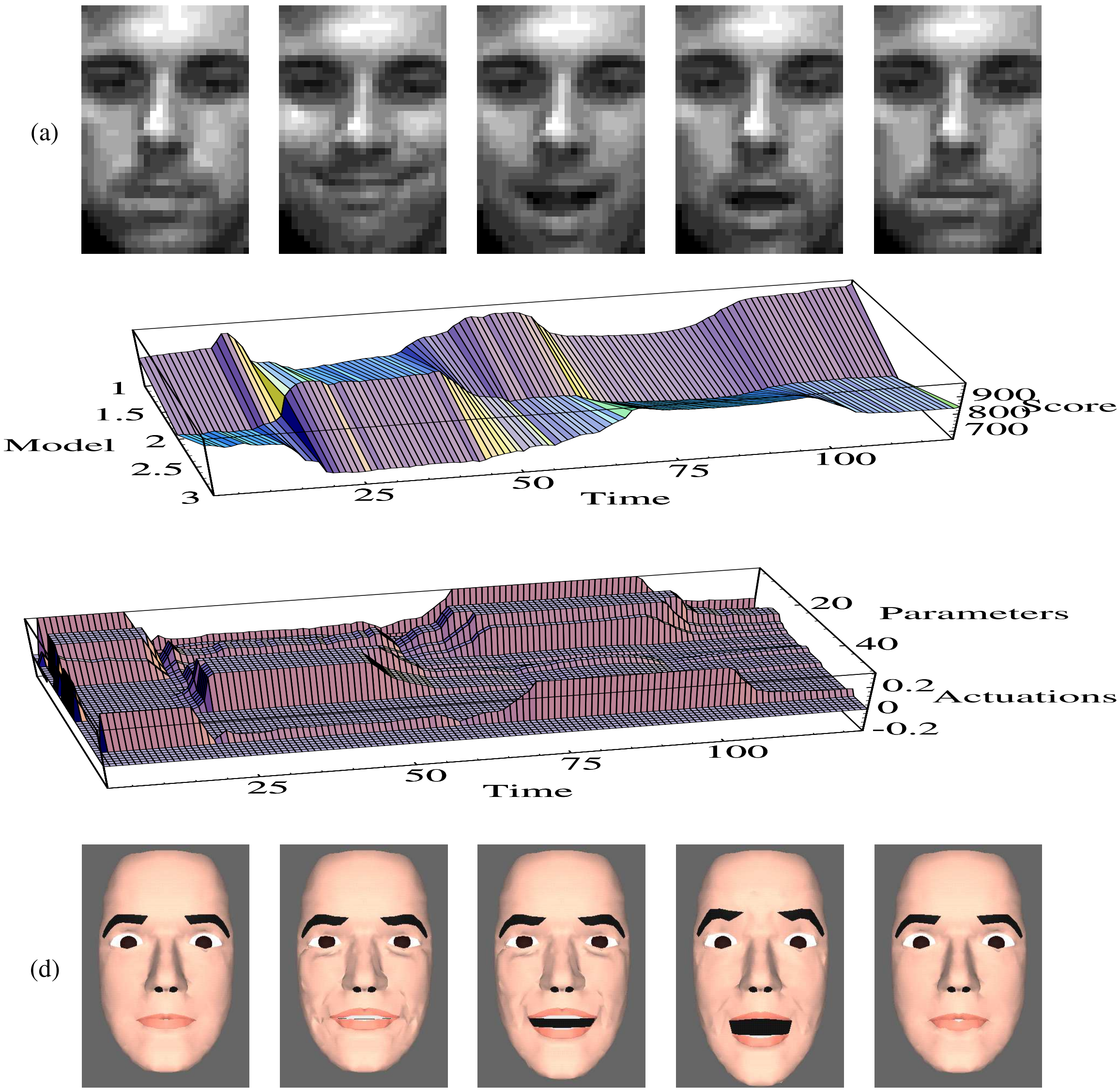


Figure 13: *(a) Face images used as input, (b) normalized correlation scores for each 2-D template, (c) resulting muscle control parameters, (d) images from the resulting facial animation.*

the live system. We have run similar experiments with local templates of the face and with a larger number of expressions.

4 Head Tracking and Orientation

One of the major constraints of the above described system is its inability to deal with large motions of the head. Tracking head positions and orientation is extremely important for both understanding facial movements and generating realistic facial and head motions. Consequently, we developed a system that can accurately track the head under virtually all conditions, including large head motions and low frame rates, independent of the same points on the head being visible over the entire length of the sequence.

Our approach is to interpret the optical flow field using a three-dimensional model. We use an ellipsoidal model of the head, which is a good approximate to the entire shape and can be automatically initialized with reasonable accuracy. The technique we use for tracking this model may be considered as *motion regularization* or *flow regularization*. The unconstrained optical flow is first computed for the entire sequence, and the rigid motion of the 3-D head model that best accounts for the observed flow is interpreted as the motion of the head. This is much in the style of Horowitz and Pentland [23]. The model's 3-D location and rotation is then modified by these parameters, and used as the starting point for interpreting the next frame, and so on (see Basu, Essa and Pentland [2] for further details). Our experiments (shown below) demonstrate that this method can provide

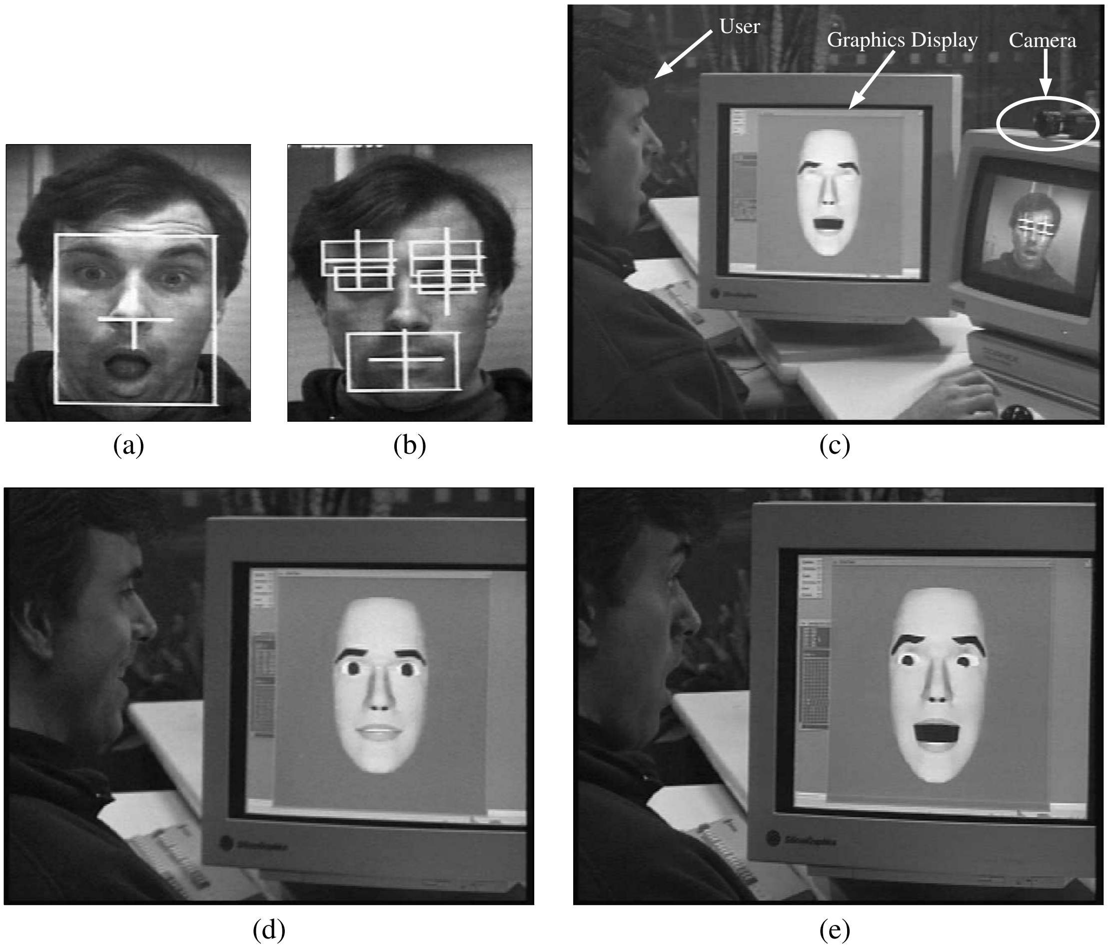


Figure 14: *(a) Face with single template, (b) Face with multiple templates. (c) Complete system tracking eyes, mouth, eyebrows., (d) tracking a smile and (e) a surprise expression.*

very robust tracking over hundreds of image frames for a very wide range of head motions.

Experiments: To demonstrate the tracking performance of this system we have presented several example sequences in the figures below. In Figure 15, several key frames from a sequence captured at 30 FPS with a Sony HandyCam are shown. The first row of images contains the original images from the sequence, while the next row shows tracking using an ellipsoidal model. The ellipsoidal model is superimposed on the image.

To demonstrate the accuracy of the system's position and orientation estimates, we have compared the results to a calibrated synthetic sequence. This sequence was generated by animating a synthetic head (model courtesy of Viewpoint Data Labs Inc. [36]) using the SGI graphics libraries. The motion parameters used to drive the model were in the same format as those estimated by the system, and were obtained from running the system on a separate image sequence (not shown). As a result, the exact rigid parameters of the model were known at every frame. The results of this experiment are shown in Figure 16 below. Again, several key frames are shown from the original sequence, followed by the tracking by the ellipsoidal model. Below these key frames, a separate plot is shown for each rigid parameter. The "model" line corresponds to the actual rigid parameters of the animated head and the "ellipsoid" line corresponds to the parameters estimated using our method.

As in the sequence shown in Figure 15, it is clear that our tracking maintains good point to point correspondence (*i.e.,* , point on the model to point on the head) over the whole sequence. We have also attempted tracking the same sequences using a 2-D planar patch and found that estimated orientations are far more accurate for the 3-D ellipsoidal model than for the 2-D planar model. The ellipsoidal model also produces slightly better estimates of the translation parameters. It is the detailed orientation information that this system extracts, though, that is its most significant advantage over other schemes. This is due to the explicit 3-D nature of the model.

5 Conclusions and Future Work

The automatic analysis and synthesis of facial expressions is becoming increasingly important in human-machine interaction. Consequently, we have developed a mathematical formulation and implemented a computer system capable of detailed analysis, tracking and synthesis of facial expressions and head movements within an active and dynamic (analysis-synthesis) framework. The purpose of this system is first to analyze real facial motion to obtain an improved computer model of facial and head movements, and then to use the improved model to create extended facial animation sequences by automatically analyzing video of real humans.

This system analyzes facial expressions by observing expressive articulations of a subject's face in video sequences. For detailed facial modeling, the visual observation (sensing) is achieved by using an optical flow method. For facial animation, the visual observation is achieved by using normalized correlation with 2-D templates. In both cases the observed motion is coupled to a physical model describing the skin and muscle structure, and the muscle control variables estimated.

We have also developed a head tracking system that can extract head positions and orientations very robustly for a large range of head motions.

Our experiments to date have demonstrated that we can indeed extract FACS-like models that are more detailed than existing models. We have also demonstrated the ability to create extended animation sequences in real time by analysis of video input, making use of the assumption that the animation consists of a limited set of facial motions.

We are now processing data from a wider range of facial expression with head motions in order to develop a model that is adequate for "all" facial expressions, and working to make the real-time animation system more person-independent. We are also improving on on our model to robustly handle finer lip and eye motions and also to deal with wrinkles and texture with much more detailed 3-D models.

Acknowledgments

We would like to thank Baback Moghaddam, Eero Simoncelli, and John Y. Wang for their help. Also thanks to Viewpoint Data Labs Inc. [36] for sharing their 3-D models.

References

[1] A. Azarbayejani, B. Horowitz, and A. Pentland. Recursive estimation of structure and motion using the relative orientation constraint. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 1993.

[2] Sumit Basu, Irfan Essa, and Alex Pentland. Motion regularization for model-based head tracking. Technical Report 362, MIT Media Laboratory, Perceptual Computing Section, January 1996. Available as MIT Media Lab Perceptual Computing Techreport # 362 from http://www-white.media.mit.edu/vismod/.

[3] M. J. Black and Y. Yacoob. Tracking and recognizing rigid and non-rigid facial motions using local parametric model of image motion. In *Proceedings of the International Conference on Computer Vision*, pages 374–381. IEEE Computer Society, Cambridge, MA, 1995.

[4] T. Darrell and A. Pentland. Space-time gestures. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 1993. Available as MIT Media Lab Perceptual Computing Techreport # 197 from http://www-white.media.mit.edu/vismod/.

[5] P. Ekman. Facial expression of emotion: An old controversy and new findings. *Philosophical Transactions: Biological Sciences (Series B)*, 335(1273):63–69, 1992.

[6] P. Ekman and W. V. Friesen. *Facial Action Coding System*. Consulting Psychologists Press Inc., 577 College Avenue, Palo Alto, California 94306, 1978.

[7] P. Ekman, T. Huang, T. Sejnowski, and J. Hager (Editors). Final Report to NSF of the Planning Workshop on Facial Expression Understanding. Technical report, National Science Foundation, Human Interaction Lab., UCSF, CA 94143, 1993.

[8] A. Emmett. Digital portfolio: Tony de peltrie. *Computer Graphics World*, 8(10):72–77, October 1985.

[9] I. Essa. *Analysis, Interpretation, and Synthesis of Facial Expressions*. PhD thesis, Massachusetts Institute of Technology, MIT Media Laboratory, Cambridge, MA 02139, USA, 1994. Available as MIT Media Lab Perceptual Computing Techreport # 303 from http://www-white.media.mit.edu/vismod/.

[10] I. Essa, T. Darrell, and A. Pentland. Tracking facial motion. In *Proceedings of the Workshop on Motion of Nonrigid and Articulated Objects*, pages 36–42. IEEE Computer Society, 1994. Available as MIT Media Lab Perceptual Computing Techreport # 272 from http://www-white.media.mit.edu/vismod/.

[11] I. Essa and A. Pentland. A vision system for observing and extracting facial action parameters. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 76–83. IEEE Computer Society, 1994. Available as MIT Media Lab Perceptual Computing Techreport # 247 from http://www-white.media.mit.edu/vismod/.

[12] S. Glenn. VActor animation system. In *ACM SIGGRAPH Visual Proceedings*, page 223, SimGraphics Engineering Corporation, 1993.

[13] B. K. P. Horn and B. G. Schunck. Determining optical flow. *Artificial Intelligence*, 17:185–203, 1981.

[14] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active contour models. *International Journal of Computer Vision*, 1(4):321–331, 1987.

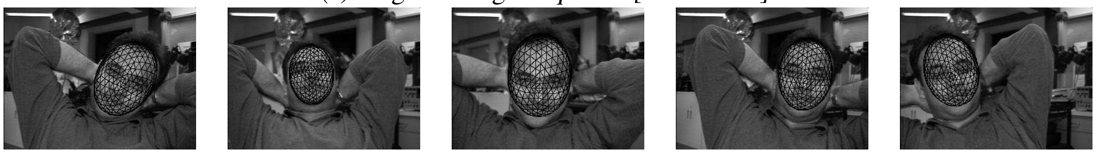
[15] Y. Lee, D. Terzopoulos, and K. Waters. Realistic modeling for facial animation. In *ACM SIGGRAPH Conference Proceedings*, 1995.

[16] P. Litwinowitcz and L. Williams. Animating images with drawings. *ACM SIGGRAPH Conference Proceedings*, pages 409–412, 1994. Annual Conference Series.

[17] N. Magnenat-Thalmann, E. Primeau, and D. Thalmann. Abstract muscle action procedures for face animation. *The Visual Computer*, 3:290–297, 1988.

Frame 33 | Frame 100 | Frame 121 | Frame 180 | Frame 256

(a) Original Image Sequence [300 frames].



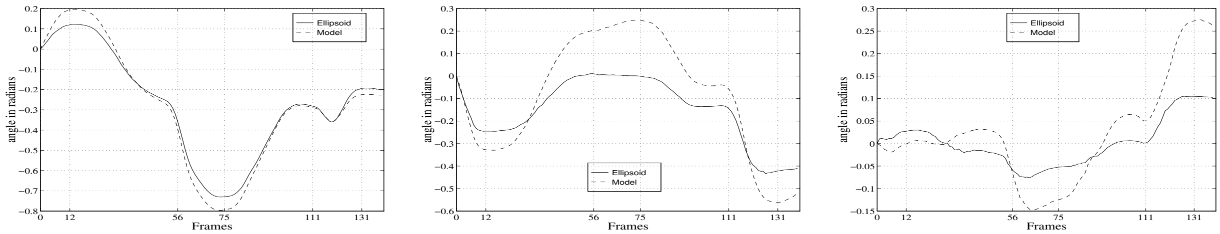
(b) Tracking using 3-D ellipsoidal model.

Figure 15: *Results of tracking on sequence acquired at 30 fps (using JPEG compression) and 320x240 resolution.*

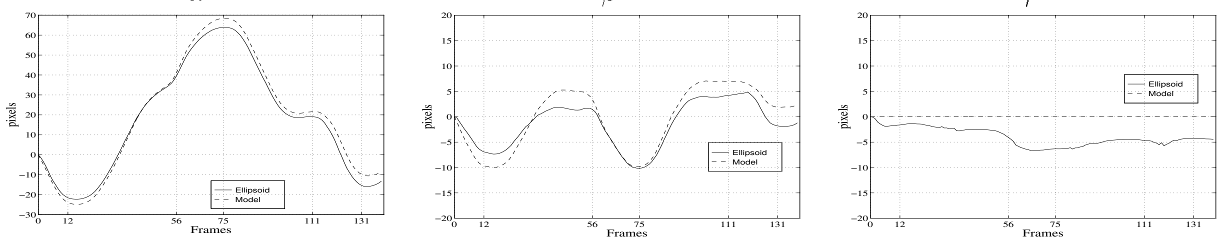
Frame 12 | Frame 56 | Frame 75 | Frame 111 | Frame 131

(a) Original Image Sequence [300 frames].

(b) Tracking using 3-D ellipsoidal model.



α β γ



t_x t_y t_z

Figure 16: *Results of tracking on a synthetic sequence. Row (a) shows the model sequence, row (b) shows our 3-D model for tracking. The plots show the comparison for the six parameters between modeled (dotted line) and our analysis (solid line). Angles are in Radians and positions in Pixels.*

[18] B. Moghaddam and A. Pentland. Face recognition using view-based and modular eigenspaces. In *Automatic Systems for the Identification and Inspection of Humans*, volume 2277. SPIE, 1994.

[19] B. Moghaddam and A. Pentland. Probabistic visual learning for object detection. In *Proceedings of the International Conference on Computer Vision*. IEEE Computer Society, 1995. Available as MIT Media Lab Perceptual Computing Techreport # 326 from http://www-white.media.mit.edu/vismod/.

[20] F. Parke. Parameterized modeling for facial animation. *IEEE Computer Graphics and Applications*, 2(9):61–68, 1982.

[21] F. I. Parke. Techniques of facial animation. In Nadia Magnenat Thalmann and Daniel Thalmann, editors, *New Trends in Animation and Visualization*, chapter 16, pages 229–241. John Wiley and Sons, 1991.

[22] C. Pelachaud, N. Badler, and M. Viaud. Final Report to NSF of the Standards for Facial Animation Workshop. Technical report, National Science Foundation, University of Pennsylvania, Philadelphia, PA 19104-6389, 1994.

[23] A. Pentland and B. Horowitz. Recovery of nonrigid motion and structure. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 13(7):730–742, July 1r991.

[24] A. Pentland, B. Moghaddam, and T. Starner. View-based and modular eigenspaces for face recognition. In *Computer Vision and Pattern Recognition Conference*, pages 84–91. IEEE Computer Society, 1994. Available as MIT Media Lab Perceptual Computing Techreport # 245 from http://www-white.media.mit.edu/vismod/.

[25] A. Pentland, B. Moghaddam, and T. Starner. View-based and modular eigenspaces for face recognition. In *Computer Vision and Pattern Recognition Conference*, pages 84–91. IEEE Computer Society, 1994.

[26] S. Pieper, J. Rosen, and D. Zeltzer. Interactive graphics for plastic surgery: A task level analysis and implementation. *Computer Graphics, Special Issue: ACM Siggraph, 1992 Symposium on Interactive 3D Graphics*, pages 127–134, 1992.

[27] S. M. Platt and N. I. Badler. Animating facial expression. *ACM SIGGRAPH Conference Proceedings*, 15(3):245–252, 1981.

[28] T. Poggio and F. Girosi. A theory of networks for approximation and learning. Technical Report A.I. Memo No. 1140, Artificial Intelligence Lab, MIT, Cambridge, MA, July 1989.

[29] M. Rydfalk. *CANDIDE: A Parameterized Face*. PhD thesis, Linköping University, Department of Electrical Engineering, Oct 1987.

[30] A. Saulnier, M. L. Viaud, and D. Geldreich. Real-time facial analysis and synthesis chain. In *International Workshop on Automatic Face and Gesture Recogntion*, pages 86–91, Zurich, Switzerland, 1995. Editor, M. Bichsel.

[31] E. P. Simoncelli. *Distributed Representation and Analysis of Visual Motion*. PhD thesis, Massachusetts Institute of Technology, 1993.

[32] D. Terzopoulos and R. Szeliski. Tracking with kalman snakes. In A. Blake and A. Yuille, editors, *Active Vision*, pages 3–20. MIT Press, 1993.

[33] D. Terzopoulos and K. Waters. Physically-based facial modeling, analysis, and animation. *The Journal of Visualization and Computer Animation*, 1:73–80, 1990.

[34] D. Terzopoulus and K. Waters. Analysis and synthesis of facial image sequences using physical and anatomical models. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 15(6):569–579, June 1993.

[35] M. Viaud and H. Yahia. Facial animation with muscle and wrinkle simulation. In *IMAGECON 1993: Second International Conference on Image Communication*, pages 117–121, 1993.

[36] Viewpoint Data Labs. 625 S. State, Orem, UT 84058, USA, +1-800-DATASET. http://www.viewpoint.com/home.shtml.

[37] S. A. Wainwright, W. D. Biggs, J. D. Curry, and J. M. Gosline. *Mechanical Design in Organisms*. Princeton University Press, 1976.

[38] J. Y. A. Wang and E. Adelson. Layered representation for motion analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 1993.

[39] K. Waters. A muscle model for animating three-dimensional facial expression. *ACM SIGGRAPH Conference Proceedings*, 21(4):17–23, 1987.

[40] K. Waters and D. Terzopoulos. Modeling and animating faces using scanned data. *The Journal of Visualization and Computer Animation*, 2:123–128, 1991.

[41] L. Williams. Performance-driven facial animation. *ACM SIGGRAPH Conference Proceedings*, 24(4):235–242, 1990.