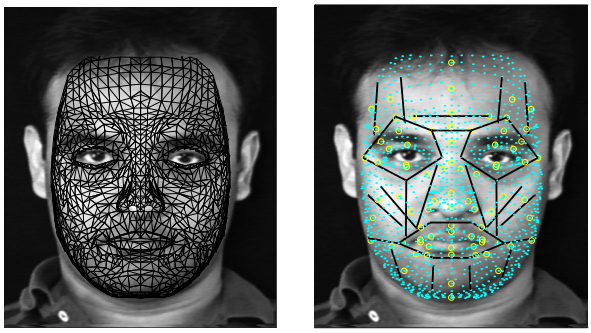
M.I.T Media Laboratory Perceptual Computing Section Technical Report No. 272
Appears: Proceedings of the IEEE Workshop on Nonrigid and Articulate Motion, Austin, Texas, November 1994

Tracking Facial Motion

Irfan A. Essa, Trevor Darrell and Alex Pentland
Perceptual Computing Section, The Media Laboratory
Massachusetts Institute of Technology
Cambridge MA 02139, U.S.A.

Abstract

We describe a computer system that allows real-time tracking of facial expressions. Sparse, fast visual measurements using 2-D templates are used to observe the face of a subject. Rather than track features on the face, the distributed response of a set of templates is used to characterize a given facial region. These measurements are coupled via a linear interpolation method to states in a physically-based model of facial animation, which includes both skin and muscle dynamics. By integrating real-time 2D image-processing with 3-D models, we obtain a system that is able to quickly track and interpret complex facial motions.



(a) Mesh (b) Muscles

Figure 1: *(a) Face image with a FEM mesh placed accurately over it and (b) Face image with muscles (black lines), and nodes (dots).*

1 Introduction

The communicative power of the face makes the modeling of facial expressions and the tracking of the expressive articulations of a face an important problem in computer vision and computer graphics. Consequently, several researchers have begun to develop methods for tracking of facial expression [10, 11, 16, 18].

These efforts, while exciting and important, have had limitations such as requiring makeup, and hand-initialization of the facial model. In this paper we improve on these previous systems by removing the need for surface markings and hand-initialization. We describe a tool for for real-time facial tracking, using spatio-temporal normalized correlation measurements [4] from video which are interpreted using a physically-based facial modeling system [7].

The principle difficulty in real-time tracking of facial articulations is the sheer complexity of human facial movement. To represent facial motion using a low-order model, many systems define independent geometric (*i.e.,* FACS [6]) and physical [5, 7, 17] parameters for modeling facial motion. The combinations of these parameters (mostly called "Action Units") results in a large set of possible facial expressions. The level of detail of facial motion encompassed by each parameter provides a broad base for representing complex facial articulations. Tracking of facial expressions can be achieved by categorizing a set of such 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. In other words, in articulation of facial expressions, there are often only few independent parameters each of which may have a large amount of temporal and spatial variation.

With this low-order model, we can use direct visual measurements to establish facial parameters in real time. In our 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; there is smooth degradation as input faces differ from the templates. We use commercial image processing hardware so that the image measurement process can occur at near-frame rate (processing 5-10 frames a second). These measurements are then coupled with our physically-based model's parameters via an interpolation process, resulting in a real-time facial tracking system.

1.1 Previous Work

There have been several attempts to track facial expressions over time. The *VActor* system [9], for instance, uses physical probes or infrared markers to measure movement of the face. Another method, which has been used to produce computer animations, is that of Williams *et al.* [18]. In this approach marks are placed on peoples faces, to track facial motion with cameras. Terzopoulos and Waters [16] developed a method to trace linear facial features, estimate corresponding parameters of a three dimensional wireframe face model, and reproduce facial expression. Requirement of facial markings for successful tracking is a significant limitation of these systems. Mase and Pentland [11, 12] introduced a method to track facial action units using optical flow. Haibo Li, Pertti Roivainen and Robert Forcheimer [10] propose a feedback control loop between vision measurements and a facial model for improved tracking. Essa and Pentland [7] describe an approach which combines optical flow with a physically-based optimal observation, estimation and control formulation to obtain better response and accuracy.

In this paper we present a method which builds upon this prior work, utilizing the power of the 3-D facial models, but turning to a fast, real-time method for facial state estimation, rather than using explicit facial landmarks (which require makeup) or dense optical flow (which is difficult to compute in real time).

2 Nonrigid Facial Modeling

To interpret and interpolate facial state estimates, we use a 3-D model of facial dynamics coupled with model of facial action units. This model captures how expressions are generated by muscle actuations and the resulting skin and tissue deformations. Hence, *a priori* information about facial structure is an important parameter for our framework. We need a model capable of controlled nonrigid deformations of various facial regions, in a fashion similar to how humans generate facial expressions by muscle actuations attached to facial tissue. For this facial model, we use a 3-D finite element mesh as shown in Figure 1(a). This is an elaboration of the mesh developed by Platt and Badler [14]. We extend this into a topologically invariant physics-based model by adding anatomically-based muscles to it (Figure 1(b)).

A physically-based dynamic model of a face, capable of articulated nonrigid deformations, requires use of Finite Element methods. These methods give our facial model an *anatomically-based* facial structure by modeling facial tissue/skin, and muscle actuators, with a geometric model to describe force-based deformations and control parameters.

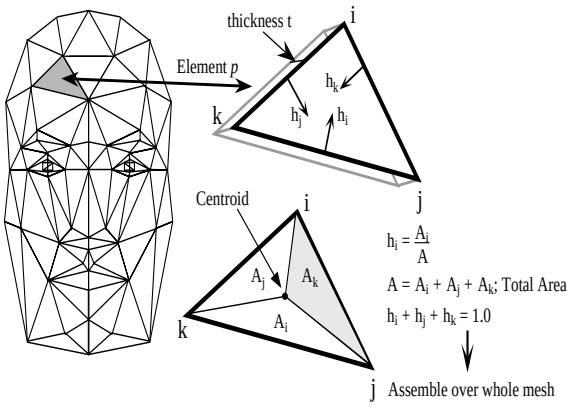


Figure 2: *Using FEM on the facial mesh to determine the continuum mechanics parameters of the skin.*

For dynamic modeling we need to integrate the system dynamics with respect to the following equation of rigid and nonrigid dynamics.

$$\mathbf{M}\ddot{\mathbf{u}} + \mathbf{D}\dot{\mathbf{u}} + \mathbf{K}\mathbf{u} = \mathbf{R}. \qquad (1)$$

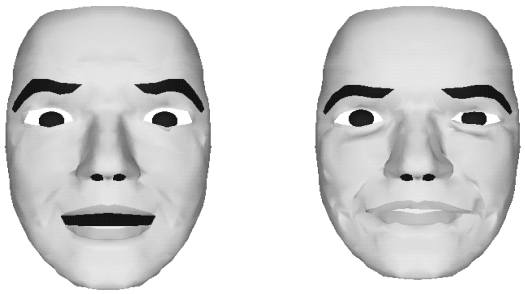
where $\mathbf{u} = [U, V, W]^T$ is the global deformation vector, which describes the deformation in the facial structure over time. Using a polygonal mesh of a face as the finite element mesh with n nodes and m elements, then $\mathbf{M}$ is a $(3n \times 3n)$ mass matrix, which accounts for the inertial properties of the face, $\mathbf{K}$ is a $(3n \times 3n)$ stiffness matrix, which accounts for the internal energy due to its elastic properties, and $\mathbf{D}$ is a $(3n \times 3n)$ damping matrix. Vector $\mathbf{R}$, is a $(3n \times 1)$ applied load vector, characterizing the force actuations of the muscles (see [8, 1] for additional details).

By defining each of the triangles on the polygonal mesh of a face as an *isoparametric triangular shell element*, (shown in Figure 2), we can calculate the mass, stiffness and damping matrices for each element (using $dV = tdA$), given the material properties of skin. Then by the assemblage process of the direct stiffness method [1, 8] on m elements, the required matrices for the whole mesh can be determined. As the integration to compute the matrices is done prior to the assemblage of matrices, each element may have different thickness t, although large differences in thickness of neighboring elements are not suitable for convergence [1]. Models for muscles are attached to this physical model of the facial tissue, based on the work of Pieper [13] and Waters [17].

2.1 Visually extracted Facial Expressions

The method of Essa and Pentland [7] provides us with a detailed physical model and also a way of observing and extracting the "action units" of a face using video sequences

(a) Surprise Image *(b) Smile Image*



(c) Surprise Model *(d) Smiling Model*

Figure 3: *Determining of expressions from video sequences. (a) and (b) show expressions of smile and surprise, (c) and (d) show a 3D model with surprise and smile expressions. This technique of extracting motor controls for facial expressions is described in [7].*

as input. The visual observation (sensing) is achieved by using an optimal estimation optical flow method coupled with a geometric and a physical (muscle) model describing the facial structure. This modeling results in a time-varying spatial patterning of facial shape and a parametric representation of the independent muscle action groups, responsible for the observed facial motions. We will use these physically-based muscle control units as our predefined facial actions in the next section. Figure 3 shows smile and surprise expressions, and the extracted smile and surprise expression on a 3D model. For further details on this method see [7].

3 Tracking of Facial Expressions

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

The number of degrees of freedom required for tracking facial articulations is limited, especially as most of the facial expressions are linear combinations of simpler motions. One can think of tracking being limited to a fixed, relatively small set of "control knobs," one for each type of motion, and then tracking the change in facial expression by moving these control knobs appropriately. The *muscle* parameters associated with these control knobs are determined by off-line modeling of each individual type of facial action as described in previous section.

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

3.1 Image Measurement

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 then interpolate the correct control parameter settings. Darrell and Pentland have successfully used this general approach to describe and recognize hand gestures [4], and in our experience this method of determining descriptive parameters is much more robust and efficient than measuring landmark positions.

By constraining the space of expressions to be recognized, we can match and recognize predefined expressions rather than having to derive new force controls for each new frame of video input. This can dramatically improve the speed of the system. Our method, therefore, begins by acquiring detailed muscle actuation and timing information for a set of expressions, using the optical flow method described in [7]. We then acquire training images of each expression for which we have obtained detailed force and timing information. This training process allows us to establish the correspondence between motor controls and image appearance.

Figure 4: *2-D Full-Face templates of neutral, smile and surprise expressions used for tracking facial expressions. See Figure 7 and Figure 8(a).*

Given a new image, we compute the peak normalized correlation score between *each* of the training views and the new data, thus producing $\mathbf{V}(t)$, a vector-valued similarity measurements at each instant. Note that the matching process can be made more efficient by limiting the search area to the neighborhood of where we last saw the eye, mouth, *etc*. Normally there is no exact match between the image and the existing expressions, so an interpolated motor observation $\mathbf{Y}(t)$ must be generated based on a weighted combination of expressions (our training examples).

In our system, we interpolate from vision scores to motor observations, using the Radial Basis Function (RBF) method [15] with linear basis functions. The details of using this interpolation method for real-time expression analysis and synthesis appear in [3].

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 5 shows the eye, brow, and mouth templates used in some of our tracking experiments, while Figure 4 shows full-face templates of neutral, smile and surprise expressions. (The normalized correlation calculation is carried out in real-time using 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 15^{\circ}$).

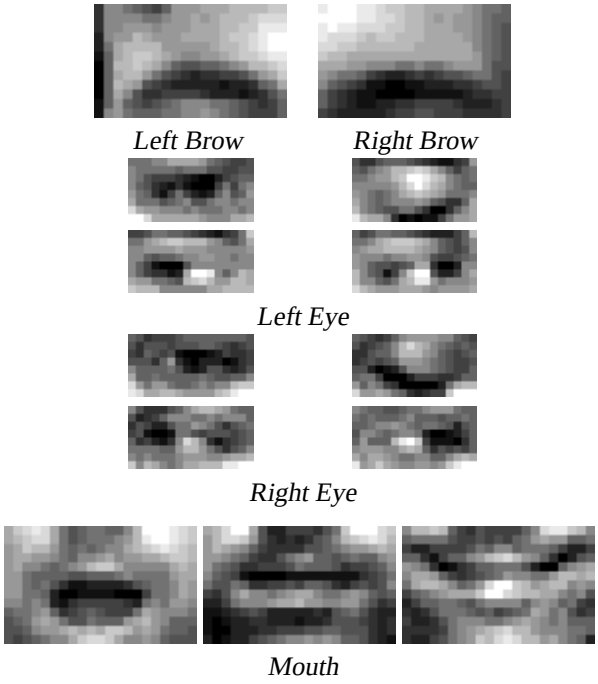


Figure 5: *2-D Eye-brows [Raised], Left and Right Eyes [Open, Closed, Looking Left, and Right], and Mouth templates [Open, Closed and Smiling] used for tracking facial expressions. See Figure 8(b).*

4 Dynamic Estimation

Estimating motor controls and then driving a physical system with the inputs from such a noisy source is prone to errors, and can result in divergence or a chaotic physical response. This is why an estimation framework needs to be incorporated to obtain stable and well-proportioned results. Similar considerations motivated the framework used in [10] or [7]. Figure 6 shows the whole framework of estimation and control of our facial expression tracking system.

This framework uses a continuous time Kalman filter (CTKF) which allows us to estimate the uncorrupted state vector, and produces an *optimal least-squares estimate* under quite general conditions [2]. The CTKF for the above system is established by the following formulation:

$$\dot{\hat{\mathbf{X}}} = \mathbf{A}\hat{\mathbf{X}} + \mathbf{L}\left(\mathbf{Y} - \mathbf{C}\hat{\mathbf{X}}\right), \qquad (2)$$

where $\hat{\mathbf{X}}$ is the linear least squares estimate of the state $\mathbf{X}$, which are the motor controls of facial motion. $\mathbf{A}$ is a state evolution matrix and contains elements of $\mathbf{K}$, $\mathbf{M}$ and $\mathbf{D}$ from Equation (1) to relate the changes in facial mesh with muscle actuation. $\mathbf{Y}$ is the observed motor state ($= \mathbf{X}$ here) for a set of correlation scores $\mathbf{V}$. Using the

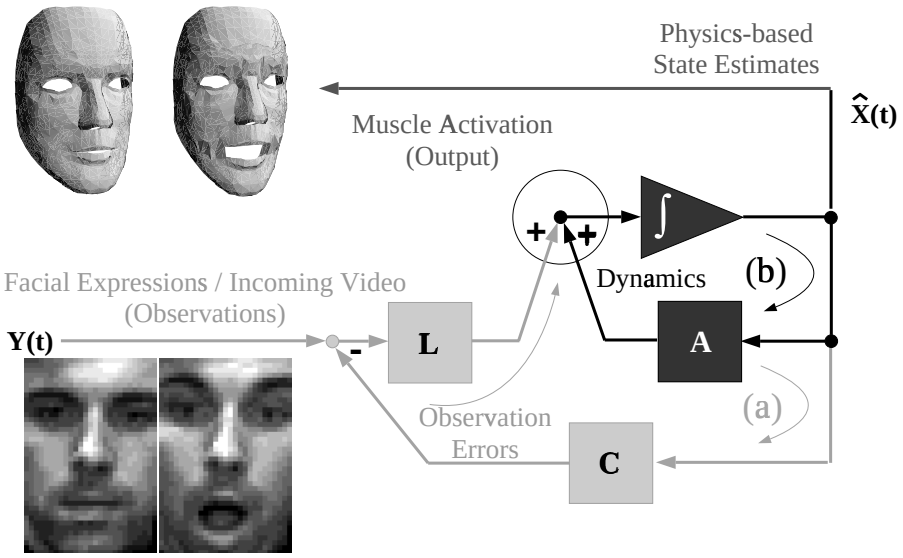


Figure 6: *Block diagram of the proposed control-theoretic approach. Showing the estimation and correction loop (a), and the dynamics loop (b).*

Riccati equation [2] to obtain the optimal error covariance matrix $\mathbf{\Lambda}_e$ with $\mathbf{\Lambda}_e$ as the error covariance matrix for $\hat{\mathbf{X}}$ and $\mathbf{\Lambda}_m$ the error covariance matrix for measurements $\mathbf{Y}$, the Kalman Gain matrix $\mathbf{L}$ is simply: $\mathbf{L} = \mathbf{\Lambda}_e \mathbf{C}^T {\mathbf{\Lambda}_m}^{-1}$.

The Kalman filter, Equation (2), mimics the noise free dynamics and corrects its estimate with a term proportional to the difference $(\mathbf{Y} - \mathbf{C}\hat{\mathbf{X}})$. This correction is between the observation and our best prediction based on previous data. Figure 6 shows the estimation loop (the bottom loop (a)) which is used to correct the dynamics based on the error predictions.

5 Experiments

Figure 7 illustrates an example of real-time facial expression tracking 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 4, and peak normalized correlation scores are measured. These scores are then converted to motor observations ($\mathbf{Y}(t)$) and fed into the muscle control loop, to produce the muscle control parameters (state estimates; $\hat{\mathbf{X}}(t)$). Five images from the resulting sequence of mimicking facial expressions in 3-D are shown in (b). This example ran in real time, with 5 frames processed per second.

Figure 8 (c), (d) and (e) show some of the live shots of the system in use. Figure 8 (a) and (b) show the video feed with the regions of interest on the face for both full-face and local region templates. We have tested this system for video sequences of upto several minutes without noticeable failure in tracking. We have also tested the system successfully for tracking lip motions for speech. The major difficulty encountered is that increasing the number of templates slows down the processing and creates a lag of about half a second to a second, which is unacceptable for some applications. We are working on reducing the lag time by incorporating a more sophisticated prediction algorithm.

6 Conclusions

The automatic analysis, synthesis and tracking 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 real-time tracking of facial expressions through extended video sequences.

This system analyzes facial expressions by observing expressive articulations of a subject's face in video sequences. The primary visual measurements are a set of peak normalized correlation scores using a set of previously-trained 2-D templates. These measurements are then coupled to a physical model describing the skin and muscle structure, and the muscle control variables estimated.

Our experiments to date have demonstrated that we can reliably track facial expressions, including independent tracking of eye and eyebrow movement, and the mouth movements involved in speech. We are currently extending

Figure 7: *(a) Face images used as input and the images of the model from the (b) resulting tracking of facial expressions.*

our system so that it can handle large head rotations, and working to remove lags in estimation/generation by use of sophisticated prediction methods.

References

[1] Klaus-Jürgen Bathe. *Finite Element Procedures in Engineering Analysis*. Prentice-Hall, 1982.

[2] Robert G. Brown. *Introduction to Random Signal Analysis and Kalman Filtering*. John Wiley & Sons Inc., 1983.

[3] Trevor Darrell, Irfan A. Essa, and Alex Pentland. Correlation and interpolation networks for real-time expression analysis/synthesis. In *Neural Information Processing Systems Conference*. NIPS, 1994. To Appear. Also available as MIT Media Lab, Perceptual Computing Section Technical Report No. 284.

[4] Trevor Darrell and Alex Pentland. Space-time gestures. In *Computer Vision and Pattern Recognition*, 1993.

[5] 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.

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

[7] Irfan A. Essa and Alex Pentland. A vision system for observing and extracting facial action parameters. In *Computer Vision and Pattern Recognition Conference*, pages 76–83. IEEE Computer Society, 1994.

[8] Irfan A. Essa, Stan Sclaroff, and Alex Pentland. Physically-based modeling for graphics and vision. In Ralph Martin, editor, *Directions in Geometric Computing*. Information Geometers, U.K., 1993.

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

[10] Haibo Li, Pertti Roivainen, and Robert Forchheimer. 3-d motion estimation in model-based facial image coding. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 15(6):545–555, June 1993.

[11] Kenji Mase. Recognition of facial expressions for optical flow. *IEICE Transactions, Special Issue on Computer Vision and its Applications*, E 74(10), 1991.

[12] Kenji Mase and Alex Pentland. Lipreading by optical flow. *Systems and Computers*, 22(6):67–76, 1991.

[13] Steven Pieper, Joseph Rosen, and David 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.

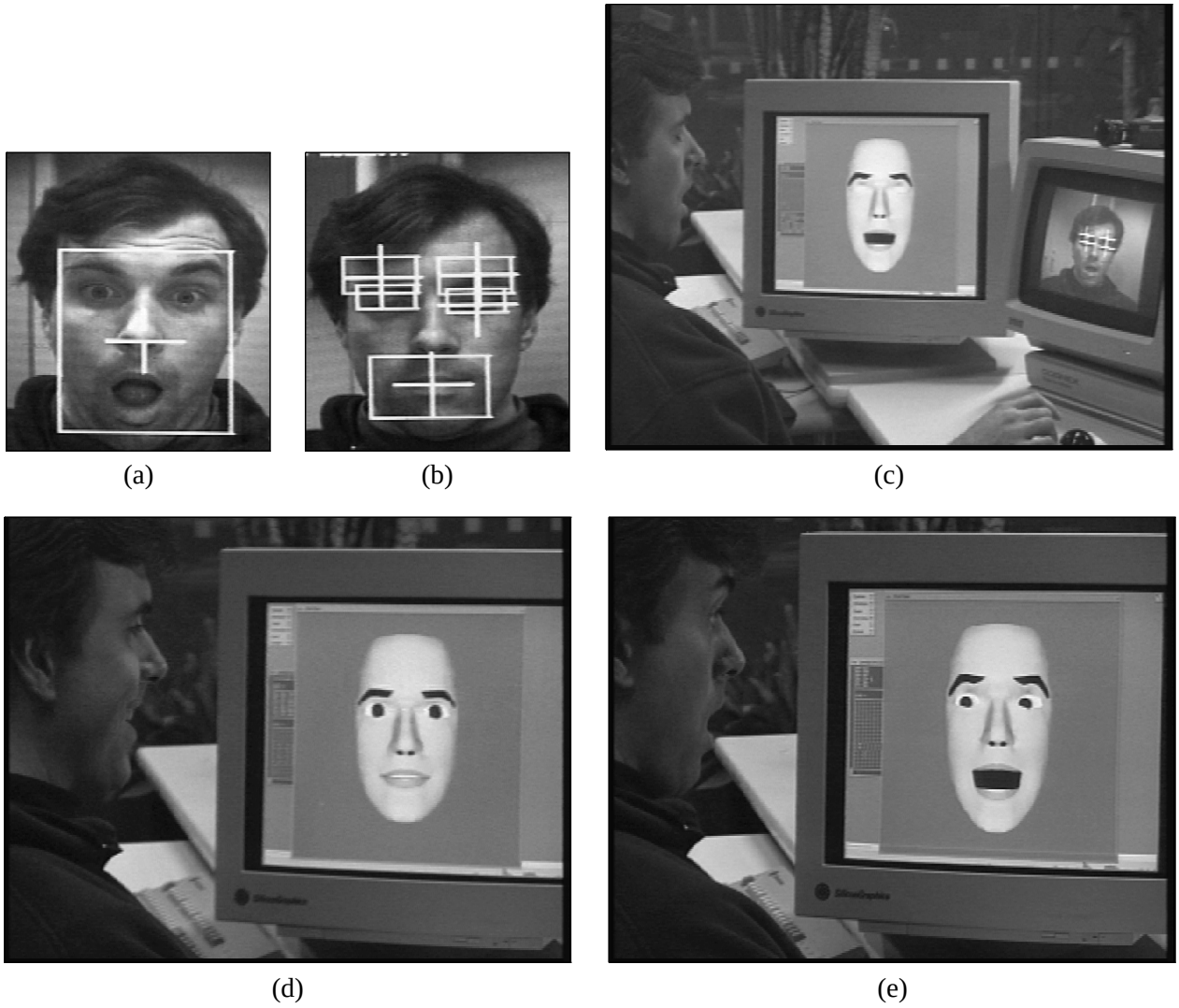


Figure 8: *(a) Face with single template, (b) Face with multiple templates. (c) Complete system tracking eyes, mouth, eyebrows., (d) tracking a smile and (e) a surprise expression.*

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

[15] 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.

[16] Demetri Terzopoulus and Keith 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.

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

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