M.I.T. Media Laboratory Perceptual Computing Section Technical Report No. 325, April 1995,
Submitted for Review (April 1995): IEEE Transactions on Pattern Analysis and Machine Intelligence

Coding, Analysis, Interpretation, and Recognition of Facial Expressions

Irfan A. Essa and Alex P. Pentland

Perceptual Computing Section, The Media Laboratory,
Massachusetts Institute of Technology
Room E15-383, 20 Ames Street, Cambridge MA 02139, U.S.A.
Telephone: (617) 253-0872, Fax:(617) 253-8874
Email: irfan@media.mit.edu, sandy@media.mit.edu

Abstract

We describe a computer vision system for observing facial motion by using an *optimal estimation* optical flow method coupled with a geometric and a physical (muscle) model describing the *facial structure*. Our method produces a reliable parametric representation of the face's independent muscle action groups, as well as an accurate estimate of facial motion.

Previous efforts at analysis of facial expression have been based on the Facial Action Coding System (FACS), a representation developed in order to allow human psychologists to code expression from static pictures. To avoid use of this heuristic coding scheme, we have used our computer vision system to probabilistically characterize facial motion and muscle activation in an experimental population, thus deriving a new, more accurate representation of human facial expressions that we call FACS+.

We use this new representation for recognition in two different ways. The first method uses the physics-based model directly, by recognizing expressions through comparison of estimated muscle activations. The second method uses the physics-based model to generate spatio-temporal motion-energy templates of the whole face for each different expression. These simple, biologically-plausible motion energy "templates" are then used for recognition. Both methods show substantially greater accuracy at expression recognition than has been previously achieved.

Categories: Facial Expressions, Expression Recognition, Face Processing, Facial Analysis, Motion and Pattern Analysis, Vision-based HCI.

1 Introduction

The communicative power of the face makes machine understanding of human expression an important problem in computer vision. There is a significant amount research on facial expressions in computer vision and computer graphics (see [9, 20] for review). Perhaps the most fundamental problem in this area is how to categorize active and spontaneous facial expressions to extract information about the underlying emotional states? [5]. Ekman and Friesen [8] have produced the most widely used system for describing visually distinguishable facial movements. This system, called the *Facial Action Coding System* or *FACS*, is based on the enumeration of all "action units" of a face which cause facial movements. As some muscles give rise to more than one action unit, the correspondence between action units and muscle units is approximate.

However, it is widely recognized that the lack of temporal and detailed spatial information (both local and global) is a significant limitation of the FACS model [9, 20]. Additionally, the heuristic "dictionary"

of facial actions originally developed for FACS-based coding of emotion has proven difficult to adapt to machine recognition of facial expression.

Instead, we would like to *objectively* quantify facial movement during various facial expressions using computer vision techniques. Consequently, the goal this paper is to provide a method for extracting an extended FACS model (FACS+), by coupling optical flow techniques with a physics-based model of both skin and muscle.

We will show that our method is capable of detailed, repeatable facial motion estimation in both time and space, with sufficient accuracy to measure previously-unquantified muscle coarticulation s in facial expressions. We will further demonstrate that the parameters extracted using this method provide improved accuracy for recognition of facial expression.

Finally, an interesting aspect of this work is that it has lead to the development of extremely simple, biologically-plausible motion energy detectors can accurately represent human expressions for coding, analysis, interpretation, tracking and synthesis of facial expressions.

1.1 Background

Representations of Facial Motion

Ekman and Friesen [8] have produced a system for describing "all visually distinguishable facial movements", called the *Facial Action Coding System* or *FACS*. It is based on the enumeration of all "action units" (AUs) of a face that cause facial movements. There are 46 AUs in FACS that account for changes in facial expression. The combination of these action units results in a large set of possible facial expressions. For example happiness expression is considered to be a combination of "pulling lip corners (AU12+13) and/or mouth opening (AU25+27) with upper lip raiser (AU10) and bit of furrow deepening (AU11)." However this is only one type of a smile; there are many variations of the above motions, each having a different intensity of actuation.

Tracking facial motion

There have been several attempts to track facial expressions over time. Mase and Pentland [16] were perhaps the first to track action units using optical flow. Although their method was simple, without a physical model and formulated statically rather than within a dynamic optimal estimation framework, the results were sufficiently good to show the usefulness of optical flow for observing facial motion.

Terzopoulos and Waters [28] developed a much more sophisticated method that tracked linear facial features to estimate corresponding parameters of a three dimensional wireframe face model, allowing them to reproduce facial expressions. A significant limitation of this system is that it requires that facial features be highlighted with make-up for successful tracking. Although active contour models (*snakes*) are used, the system is still passive; the facial structure is passively shaped by the tracked contour features without any active control based on observations.

Haibo Li, Pertti Roivainen and Robert Forchheimer [14] describe an approach in which a control feedback loop between computer graphics and computer vision processes is used for a facial image coding system. Their work is the most similar to ours, but both our goals and implementations and the details inherent in them differ. The main limitation of their work is the lack of detail in motion estimation as only large, predefined areas were observed, and only affine motion computed within each area. These limits may be an acceptable loss of quality for image coding applications. However, for our purposes this limitation is

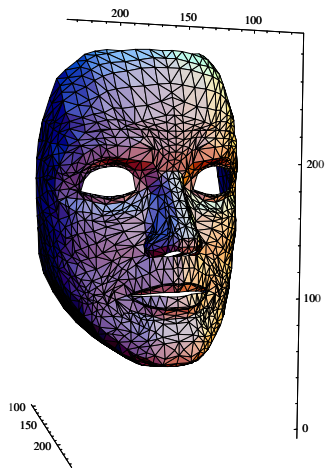


Figure 1: *Geometric Model of a Face (Polygons/Vertices).*

severe; it means we cannot observe the "true" pattern of muscle actuation because the method assumes the FACS model as the underlying representation.

Recognition of Facial Motion

Recognition of facial expressions can be achieved by categorizing a set of such predetermined facial motions as in FACS, rather than determining the motion of each facial point independently. This is the approach taken by Mase [15] and by Yacoob and Davis [32, 25] for their recognition systems. For instance Yacoob and Davis [32], following the work of Mase [15], detect motion (quantized to eight directions) in six predefined and hand initialized rectangular regions on a face, and then use a simplification of the FACS rules for the six universal expressions for recognition.

They have shown a 80% overall accuracy [1] in correctly recognizing expressions over their database of 105 expressions. Mase [15] on a smaller set of data (30 test cases) also obtained an accuracy of 80%. In many ways these are impressive results, considering the complexity of the FACS model and the difficulty in measuring facial motion within small windowed regions of the face.

In our view perhaps the principle difficulty these researchers have encountered is the sheer complexity of describing human facial movement using FACS. Using the FACS representation, there are a very large number of AUs, which combine in extremely complex ways to give rise to expressions. Moreover, there is now a growing body of psychological research that argues that it is the dynamics of the expression, rather than detailed spatial deformations, that is important in expression recognition [1, 2, 5, 6]. Indeed several famous researchers have claimed that the timing of expressions, something that is completely missing from FACS, is a critical parameter in recognizing emotions [7, 18]. To us this strongly suggests moving away from a static, "dissect-every-change" analysis of expression (which is how the FACS model was developed), towards a whole-face analysis of facial dynamics in motion sequences.

[1] This number extracted from Yacoob and Davis [32] Table 3, by dividing the correctly recognized expressions by the number of all expressions ($\frac{total-missed-confused}{total} = \frac{105-11-10}{105}$).

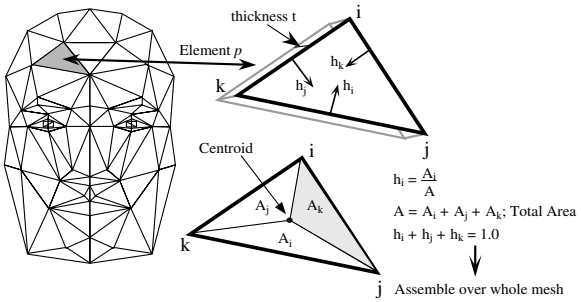


Figure 2: *Using the FACS mesh to determine the continuum mechanics parameters of the skin using FEM.*

2 Visual Coding of Facial Motion

2.1 Vision-based Sensing: Visual Motion

We use optical flow processing as the basis for perception and measurement of facial motion. We use Simoncelli's [27] method for optical flow computation, which uses a multi-scale, coarse-to-fine, Kalman filtering-based algorithm that provides good motion estimates and error-covariance information. Using this method we compute the estimated mean velocity vector $\hat{\mathbf{v}}_i(t)$, which is the estimated flow from time t to $t+1$. We also store the flow covariances $\boldsymbol{\Lambda}_{\mathbf{v}}$ between different frames for determining confidence measures and for error corrections in observations for the dynamic model (see Section 2.3 and Figure 4 [observation loop (a)]).

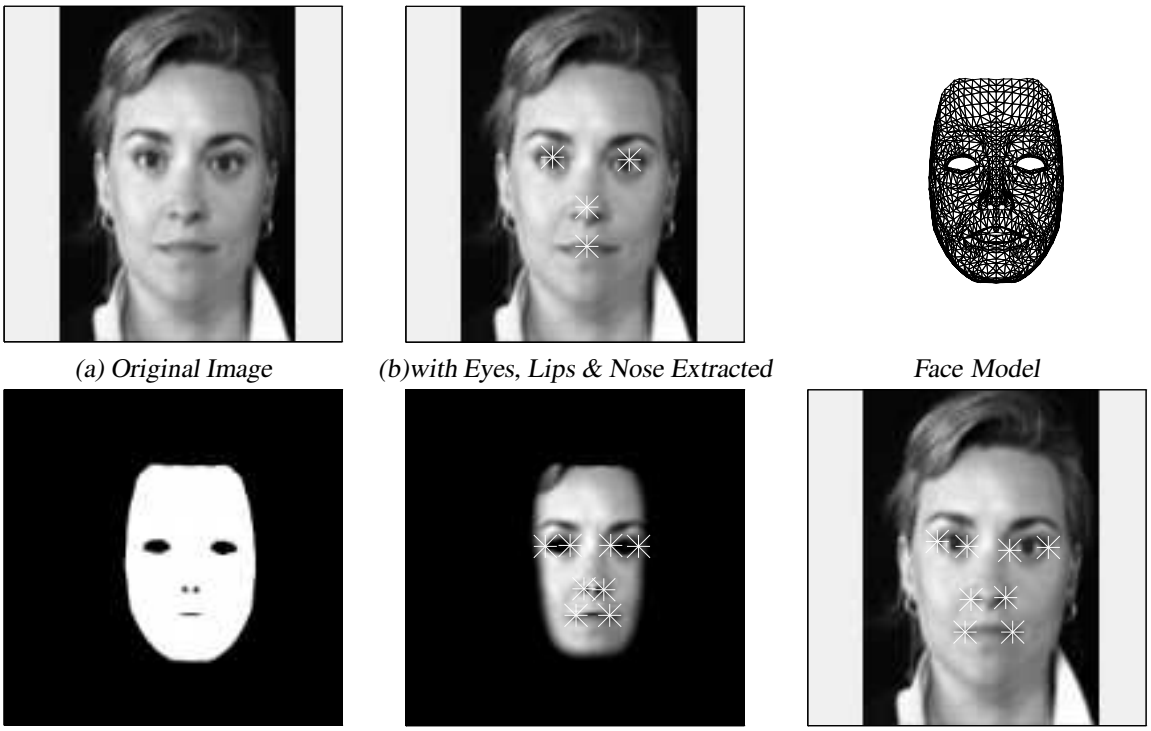
2.2 Facial Modeling

A priori information about facial structure is an important parameter for our framework. Our face model is shown in Figure 1. This is an elaboration of the mesh developed by Platt and Badler [24]. We extend this into a topologically invariant physics-based model by adding anatomically-based muscles to it [10].

In order to conduct analysis of facial expressions and to define a new suitable set of control parameters (FACS+) using vision-based observations, we require a model with time dependent *states* and *state evolution* relationships. FACS and the related AU descriptions are purely static and passive, and therefore the association of a FACS descriptor with a dynamic muscle is inherently inconsistent. This problems motivated Waters [31] to develop a muscle model in a dynamic framework. By modeling the elastic nature of facial skin and the anatomical nature of facial muscles he developed a dynamic model of the face, including FACS-like control parameters. By implementing a procedure similar to that of Waters', we also built a dynamic muscle-based model of a face.

A physically-based dynamic model of a face may be constructed by 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(see [3, 12, 17]).

By defining each of the triangles on the polygonal mesh in Figure 1 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 (acquired from [23, 29]). Then by the assemblage process of the direct stiffness method [3, 12] the required matrices for the whole mesh can be determined.



(a) Original Image (b)with Eyes, Lips & Nose Extracted Face Model

(d) Mask of Model (e) Warped & Masked with Features (f) Canonical Points Extracted

Figure 3: *Initialization on a face image using methods described by Pentland et al. [21, 19], using a canonical model of a face.*

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 [3].

The next step in formulating this dynamic model of the face is the combination of the skin model with a dynamic muscle model. This requires information about the attachment points of the muscles to the face, or in our geometric case the attachment to the vertices of the geometric surface/mesh. The work of Pieper [23] and Waters [31] provides us with the required detailed information about muscles and muscle attachments.

2.3 Dynamic Modeling and Estimation

Initialization of Model on an image

In developing a representation of facial motion and then using it to compare to new data we need to locate a face and the facial features in the image followed by a registration of these features for all faces in the database. Previously we initialized our estimation process by manually translating, rotating and deforming our 3-D facial model to fit a face in an image (see Figure 5). To automate this process we are now using the View-based and Modular Eigenspace methods of Pentland and Moghaddam [19, 21].

Using this method we can automatically extract the positions of the eyes, nose and lips in an image

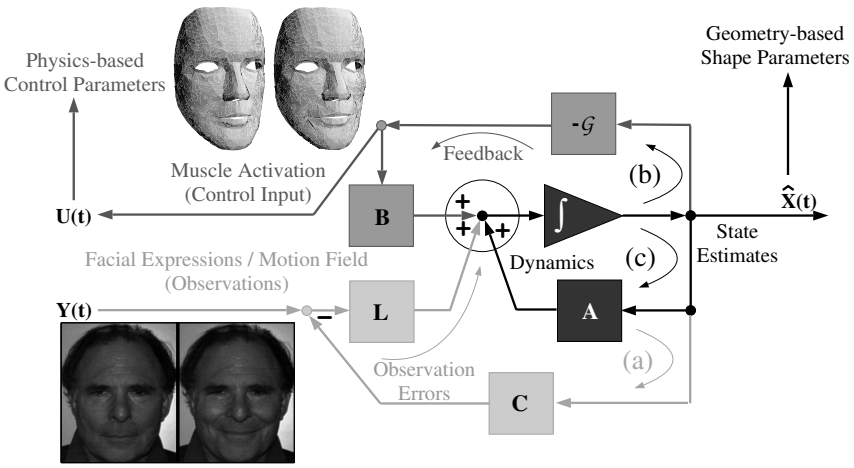


Figure 4: *Block diagram of the control-theoretic approach. Showing the estimation and correction loop (a), the dynamics loop (b), and the feedback loop (c).*

as shown in Figure 3 (b). These feature positions are used to warp the face image to match the canonical face mesh (Figure 3 (c) and (d)). This allows us to extract the additional "canonical feature points" on the image that correspond to the fixed (nonrigid) nodes on our mesh (Figure 3 (f)). After the initial registering of the model to the image the coarse-to-fine flow computation methods presented by Simoncelli [27] and Wang [30] are used to compute the flow. 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 rigid motion of the face during an expression.

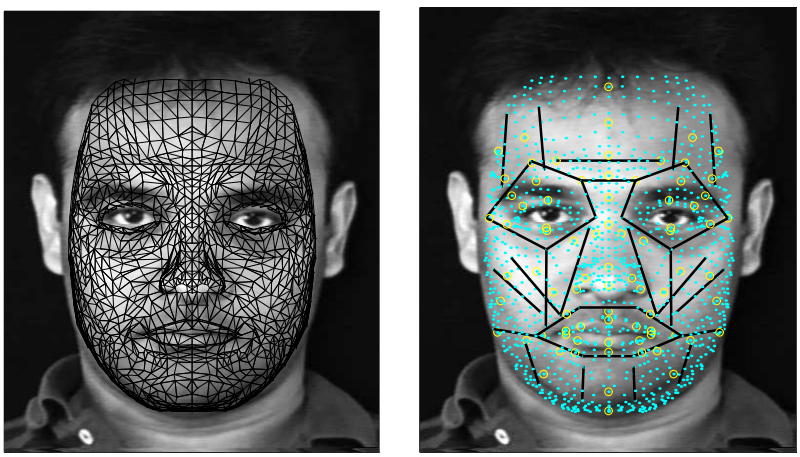
Images to face model

Simoncelli's [27] coarse-to-fine algorithm for optical flow computations provides us with an estimated flow vector, $\hat{\mathbf{v}}_i$. Now using the a mapping function, $\boldsymbol{\mathcal{M}}$, we would like to compute velocities for the vertices of the face model $\mathbf{v}_g$. Then, using the physically-based modeling techniques of Section 2.2 and the relevant geometric and physical models, we can calculate the forces that caused the motion. Since we are mapping global information from an image (over the whole image) to a geometric model, we have to concern ourselves with translations (vector $\boldsymbol{\mathcal{T}}$), and rotations (matrix $\boldsymbol{\mathcal{R}}$). The Galerkin polynomial interpolation function $\mathbf{H}$ and the strain-displacement function $\boldsymbol{\mathcal{B}}$, used to define the mass, stiffness and damping matrices on the basis of the finite element method are applied to describe the deformable behavior of the model [12, 22, 3].

We would like to use only a frontal view to determine facial motion and model expressions, and this is only possible if we are prepared to estimate the velocities and motions in the third axis (going into the image, the z-axis). To accomplish this, we define a function that does a spherical mapping, $\boldsymbol{\mathcal{S}}(u, v)$, where are u and v are the spherical coordinates. The spherical function is computed by use of a prototype 3-D model of a face with a spherical parameterization; this canonical face model is then used to wrap the image onto the shape. In this manner, we determine the mapping equation:

$$\mathbf{v}_g(x, y, z) = \boldsymbol{\mathcal{M}}(x, y, z)\hat{\mathbf{v}}_i(x, y \mid z, y) \approx \mathbf{H}\boldsymbol{\mathcal{S}}\boldsymbol{\mathcal{R}}\left(\hat{\mathbf{v}}_i(x, y) + \boldsymbol{\mathcal{T}}\right). \quad (1)$$

For the rest of the paper, unless otherwise specified, whenever we talk about velocities we will assume that the above mapping has already been applied.



(a) Mesh (b) Muscles

Figure 5: *(a) Face image with a FEM mesh placed accurately over it and (b) Face image with muscles (black lines), and nodes (dots).*

Estimation and Control

Driving a physical system with the inputs from noisy motion estimates can result in divergence or a chaotic physical response. This is why an estimation and control framework needs to be incorporated to obtain stable and well-proportioned results. Similar considerations motivated the control framework used in [14]. Figure 4 shows the whole framework of estimation and control of our active facial expression modeling system. The next few sections discuss these formulations.

The continuous time Kalman filter (CTKF) allows us to estimate the uncorrupted state vector, and produces an *optimal least-squares estimate* under quite general conditions [4, 13]. The Kalman filter is particularly well-suited to this application because it is a recursive estimation technique, and so does not introduce any delays into the system (keeping the system active). The CTKF for the above system is:

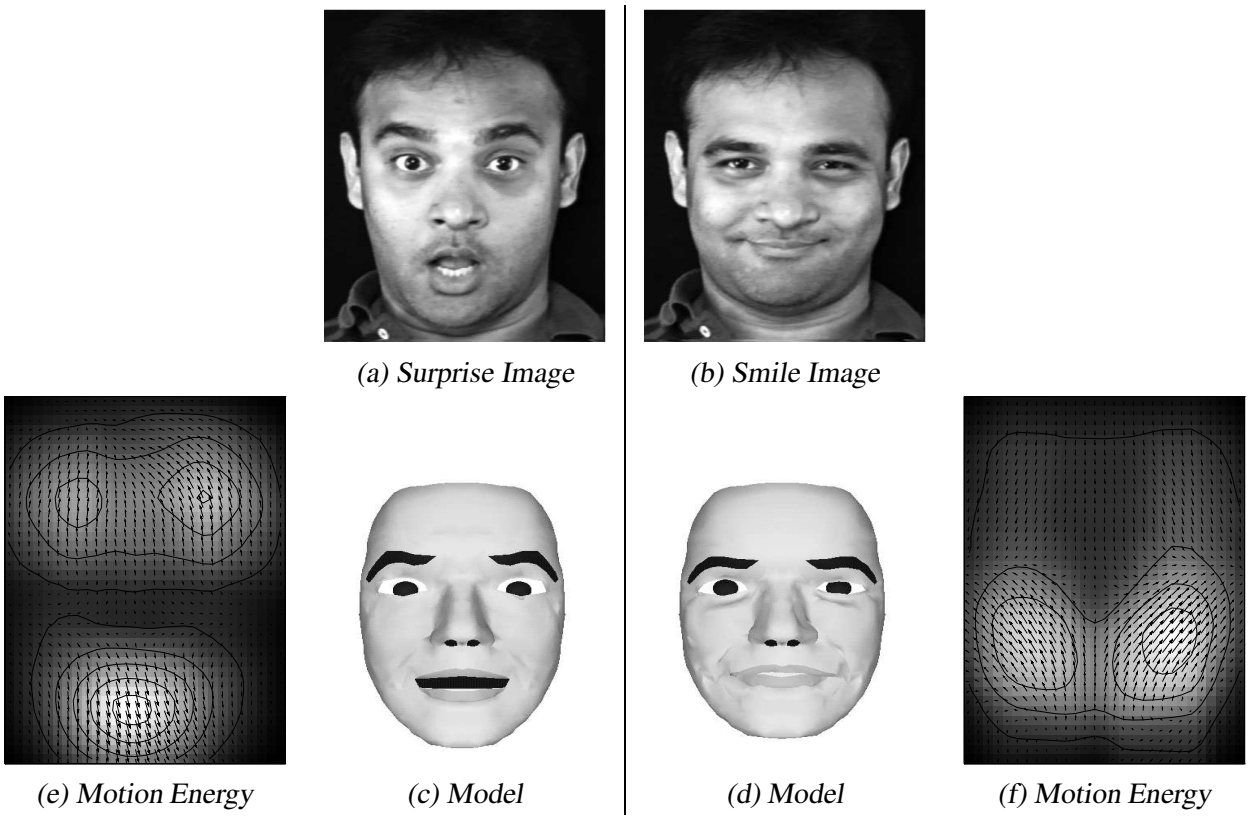
$$\dot{\hat{\mathbf{X}}} = \mathbf{A}\hat{\mathbf{X}} + \mathbf{B}\mathbf{U} + \mathbf{L}\left(\mathbf{Y} - \mathbf{C}\hat{\mathbf{X}}\right), \quad \text{where:} \quad \mathbf{L} = \mathit{\Lambda}_e \mathbf{C}^T {\mathit{\Lambda}_m}^{-1}, \tag{2}$$

where $\hat{\mathbf{X}}$ is the linear least squares estimate of $\mathbf{X}$ based on $Y(\tau)$ for $\tau < t$ and $\mathit{\Lambda}_e$ the error covariance matrix for $\hat{\mathbf{X}}$. The Kalman gain matrix $\mathbf{L}$ is obtained by solving the following Riccati equation to obtain the optimal error covariance matrix $\mathit{\Lambda}_e$:

$$\frac{d}{dt}\mathit{\Lambda}_e = \mathbf{A}\mathit{\Lambda}_e + \mathit{\Lambda}_e\mathbf{A}^T + \mathbf{G}\mathit{\Lambda}_p\mathbf{G}^T - \mathit{\Lambda}_e\mathbf{C}^T{\mathit{\Lambda}_m}^{-1}\mathbf{C}\mathit{\Lambda}_e. \tag{3}$$

The Kalman filter, Equation (2), mimics the noise free dynamics and corrects its estimate with a term proportional to the difference $(\mathbf{Y} - \mathbf{C}\hat{\mathbf{X}})$, which is the innovations process. This correction is between the observation and our best prediction based on previous data. Figure 4 shows the estimation loop (the bottom loop) which is used to correct the dynamics based on the error predictions.

The optical flow computation method has already established a probability distribution ($\mathit{\Lambda}_{\mathbf{v}}(t)$) with respect to the observations. We can simply use this distribution in our dynamic observations relationships.



(a) Surprise Image (b) Smile Image

(e) Motion Energy (c) Model (d) Model (f) Motion Energy

Figure 6: *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, and (e) and (f) show the spatio-temporal motion energy representation of facial motion for these expressions.*

Hence we obtain:

$$\mathbf{\Lambda}_m(t) = \boldsymbol{\mathcal{M}}(x, y, z)\mathbf{\Lambda}_{\mathbf{v}}(t), \quad \mathbf{Y}(t) = \boldsymbol{\mathcal{M}}(x, y, z)\hat{\mathbf{v}}_i(t). \tag{4}$$

Control, Measurement and Correction of Dynamic Motion

Now using a control theory approach we will obtain the muscle actuations. These actuations are derived from the observed image velocities. The control input vector $\mathbf{U}$ is therefore provided by the control feedback law: $\mathbf{U} = -\boldsymbol{\mathcal{G}}\mathbf{X}$, where $\boldsymbol{\mathcal{G}}$ is the *control feedback gain matrix*. We assume the instance of control under study falls into the category of an *optimal regulator* [13]. Hence, the optimal control law $\mathbf{U}^*$ is given by:

$$\mathbf{U}^* = -\mathbf{R}^{-1}\mathbf{B}^T\mathbf{P}_c\mathbf{X}^* \tag{5}$$

where $\mathbf{X}^*$ is the optimal state trajectory and $\mathbf{P}_c$ is given by solving yet another *matrix Riccati equation* [13]. Here $\mathbf{Q}$ is a real, symmetric, positive semi-definite *state weighting* matrix and $\mathbf{R}$ is a real, symmetric, positive definite *control weighting* matrix. Comparing with the control feedback law we obtain $\boldsymbol{\mathcal{G}} = \mathbf{R}^{-1}\mathbf{B}^T\mathbf{P}_c$ This control loop is also shown in the block diagram in Figure 4 (upper loop (c)).

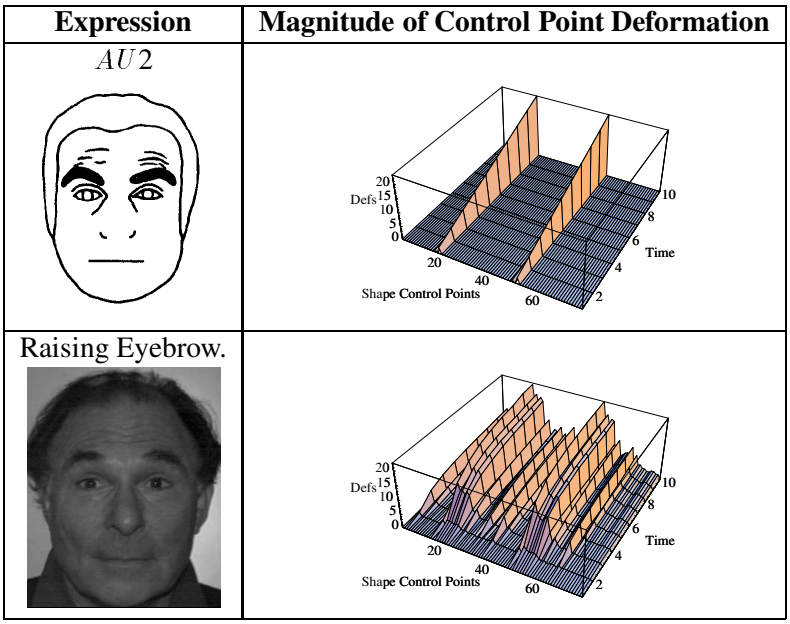


Figure 7: *FACS/*CANDIDE *deformation vs. Observed deformation for the Raising Eyebrow expression. Surface plots (top) show deformation over time for FACS actions $AU2$, and (bottom) for an actual video sequence of raising eyebrows.*

3 Analysis and Representations

The goal of this work is to develop a new representation of facial action that more accurately captures the characteristics of facial motion, so that we can employ them in recognition, coding and interpretation of facial motion. The current state-of-the-art for facial descriptions (either FACS itself or muscle-control versions of FACS) has two major weaknesses:

- The action units are purely local spatial patterns. Real facial motion is almost never completely localized; Ekman himself has described some of these action units as an "unnatural" type of facial movement. Detecting a unique set of action units for a specific facial expression is not guaranteed.
- 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. Coarticulation effects are not accounted for in any facial movement.

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. Although the muscle-based models used in computer graphics have alleviated some of these problems [31], they are still too simple to accurately describe real facial motion.

Our method lets us characterize the functional form of the actuation profile, and lets us determine a basis set of "action units" that better describes the spatial properties of real facial motion. A similar form of

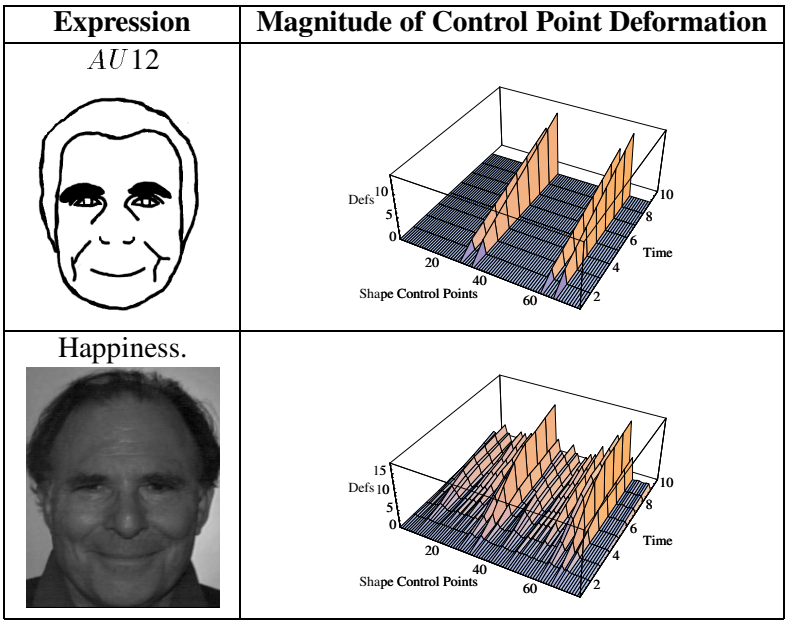


Figure 8: *FACS/*CANDIDE *deformation vs. Observed deformation for the Happiness expression. Surface plots (top) show deformation over time for FACS action* $AU\,12$*, and (bottom) for an actual video sequence of happiness.*

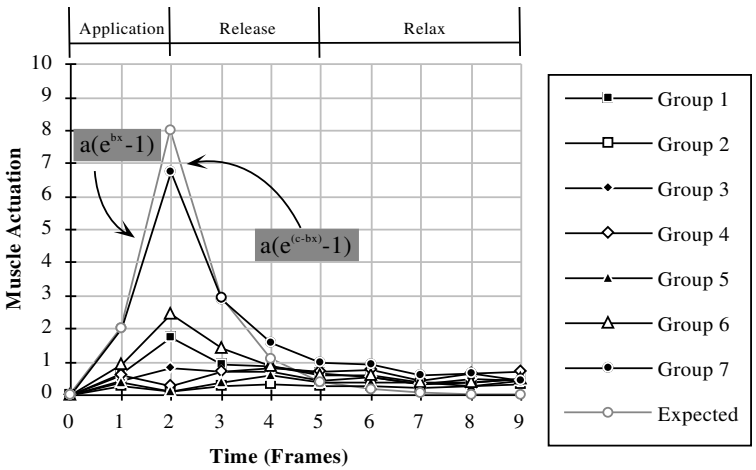


Figure 9: *Actuations over time of the seven main muscle groups for the expressions of raising brow. The plots shows actuations over time for the seven muscle groups and the expected profile of application, release and relax phases of muscle activation.*

analysis has recently been undertaken by Terzopoulos and Waters [28], where they extract muscle actuations using snakes. However their method, because it is based on sparse features, does not extract the kind of

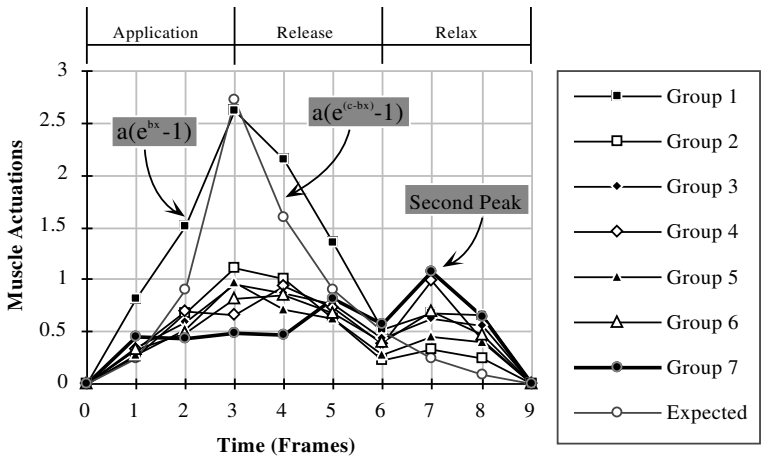


Figure 10: *Actuations over time of the seven main muscle groups for the expressions of smiling – lip motion. The plots shows actuations over time for the seven muscle groups and the expected profile of application, release and relax phases of muscle activation.*

detail that is essential for the coding, interpretation and recognition tasks we are interested in.

In the next few paragraphs, we will illustrate the resolution of our representation using the smile and eyebrow raising expressions. Questions of repeatability and accuracy will be addressed while discussing the data obtained during our expression recognition experiments.

3.1 Spatial Patterning

To illustrate that our new 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 7 and Figure 8.

The top row of Figure 7 shows $AU2$ ("Raising Eyebrow") from the FACS model and the linear actuation profile of the corresponding geometric control points. This is the type of spatio-temporal patterning commonly used in computer graphics animation. The bottom row of Figure 7 shows the observed motion of these control points for the expression of *raising eyebrow* by Paul Ekman. This plot was achieved by mapping the motion onto the FACS model and the actuations of the control points measured. As can be seen, the observed pattern of deformation is very different than that assumed in the standard implementation of FACS. There is a wide distribution of motion through all the control points, not just around the largest activation points.

Similar plots for happiness expression are shown in Figure 8. These observed distributed patterns of motion provide a detailed representation of facial motion that we will show is sufficient for accurate characterization of facial expressions.

3.2 Temporal Patterning

Another important observation about facial motion that is apparent in Figure 7 and Figure 8 is that the facial motion is far from linear in time. This observation becomes much more important when facial motion is

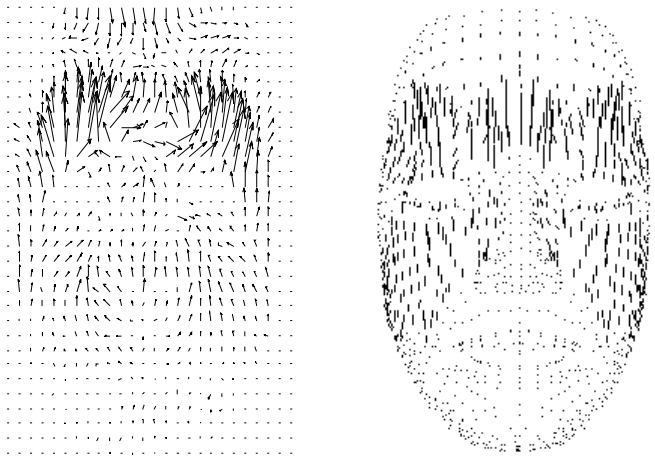


Figure 11: *Left figure shows a motion field for the expression of raise eye brow from optical flow computation and the right figure shows the motion field after it has been mapped to a face model using the control-theoretic approach of Figure 4.*

studied with reference to muscles, which is in fact the effector of facial motion and the underlying parameter for differentiating facial movements using FACS.

The top rows of Figure 7 and Figure 8, that show the development of FACS expressions can only be represented by a muscle actuation that has a step-function profile. Figure 9 and Figure 10 show plots of facial muscle actuations for the observed smile and eyebrow raising expressions. For the purpose of illustration, in this figure the 36 face muscles were combined into seven local groups on the basis of their proximity to each other and to the regions they effected. 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 for the smile expression also shows a second, delayed actuation of muscle group 7, about 3 frames after the peak of muscle group 1. Muscle group 7 includes all the muscles around the eyes and as can be seen in Figure 9 is the primary muscle group for the raising eye brow expression. This example illustrates that coarticulation effects can be observed by our system, and that they occur even in quite simple expressions. By using these observed temporal patterns of muscle activation, rather than simple linear ramps, or heuristic approaches of the representing temporal changes, we can more accurately characterize facial expressions.

3.3 Motion Templates from the Facial Model

So far we have discussed how we can extract the muscle actuations of an observed expression. However our control-theoretic approach can also be used to extract the "corrected" or "noise-free" 2-D motion field that is associated with each facial expression.

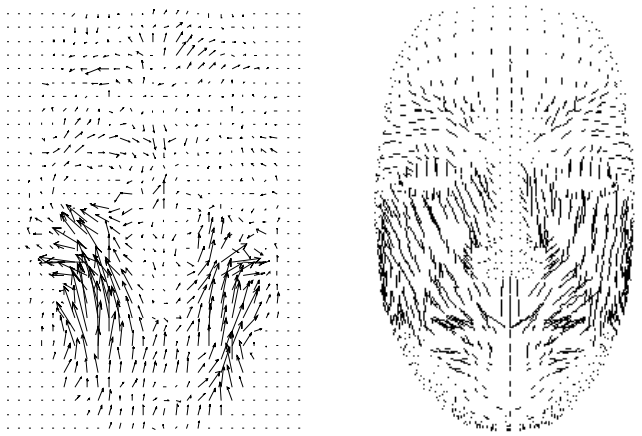


Figure 12: *Left figure shows a motion field for the expression of smile from optical flow computation and the right figure shows the motion field after it has been mapped to a face model using the control-theoretic approach of Figure 4.*

The system shown in Figure 4 employs optimal estimation, within an optimal control and feedback framework. It maps 2-D motion observations from images onto a physics-based dynamic model, and then the estimates of corrected 2-D motions (based on the optimal dynamic model) are used to correct the observations model. Figure 11 and Figure 12 show the corrected flow for the expressions of raise eyebrow and smile, and also show the corrected flow after it has been applied to the face model as deformation of the skin.

By using this methodology to *back-project* the facial motion estimates into the image we can remove the complexity of physics-based modeling from our representation of facial motion, and instead use only the 2-D observations (e.g., motion energy) to describe the facial motion that is characteristic of each facial expression.

Note that this corrected motion representation is better than could be obtained by measuring optical flow using standard optical flow techniques, because it incorporates the constraints from our physical model of the face. Figure 6 (e) and (f) shows examples of this representation of facial motion energy. It is this representation of facial motion that we will use for generating spatio-temporal "templates" for coding, interpretation and recognition of facial expressions.

4 Characterization of Facial Expressions

One of the main advantages of the methods presented here is the ability to use real imagery to define representations for different expressions. As we discussed in the last section, we do not want to rely on pre-existing models of facial expression as they are generally not well suited to our interests and needs. We would rather observe subjects making expressions and use the measured motion, either muscle actuations or 2-D motion energy, to accurately characterize each expression.

For this purpose we have developed a video database of people making expressions. Currently these subjects are video-taped while making an expression on demand. These "on demand" expressions have the

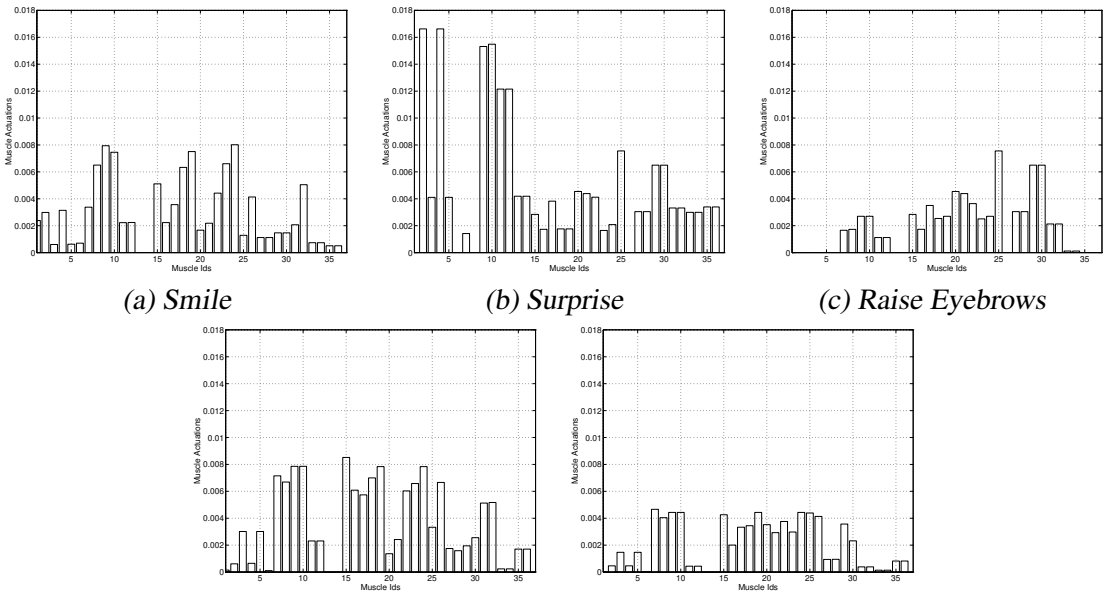
limitation that the subjects' emotion generally does not relate to his/her expression. However we are for the moment more interested in characterizing facial motion and not human emotion. Categorization of human emotion on the basis of facial expression is an important topic of research in psychology and we believe that our methods can be useful in this area. We are at present in collaborating with several psychologists on this problem.

Figure 13: *Expressions from video sequences for various people in our database. These expressions are captured at 30 frames per second at NTSC resolution and cropped to 450x380.*

At present we have a database of 20 people making expressions of *smile*, *surprise*, *anger*, *disgust*, *raise brow*, and *sad*. Some of our subjects had problems making the expression of sad, therefore we have decided to exclude that expression from our present study. We are working on expanding our database to cover many other expressions and also expressions with speech. The last frames of some of the expressions in our database are shown in Figure 13. All of these expressions are digitized as sequences at 30 frames per second and stored at the resolution of 450x380. All the results discussed in this paper are based on expressions performed by 7 people with a total of 52 expressions. This database is substantially larger than that used by Mase [15] in his pioneering work on recognizing facial expressions. Yacoob and Davis [32] currently have the largest database (30 subjects and 105 expressions). Although our database is smaller than that of Yacoob and Davis, we believe that it is sufficiently large to demonstrate that we have achieved improved accuracy at facial expression recognition.

4.1 Model-based Recognition

Recognition requires a unique "feature vector" to define each expression and a similarity metric to measure the differences between expressions. Since both temporal and spatial characteristics are important we



(a) Smile (b) Surprise (c) Raise Eyebrows

(d) Anger (c) Disgust

Figure 14: *Feature vectors of muscle templates for different expressions*

require a feature vector that can account for both of these characteristics. We must also, however, account for the speed at which the expressions are performed. Since facial expressions occur in three distinct phases: *application*, *release* and *relaxation* (see Figure 9 and Figure 10), by dividing the data into these phases and by warping it for all expressions into a fixed time period of ten discrete samples, we can normalize the temporal time course of the expression. This normalization allows us to use the muscle actuation profiles to define a unique feature vector for a each facial motion.[2] Consequently, we have defined the peak actuation of each muscle between the application and release phases as the feature vector for each expression.

We defined an *standard* muscle activation feature vector for each of the expressions *smile*, *surprise*, *anger*, *disgust*, and *raise eyebrow* by choosing at random two subjects from our database of facial expressions. These standard peak muscle actuation features, which we call *muscle templates* are shown in Figure 14. As can be seen, the muscle templates for each expression are unique, indicating that they are good features for recognition. These feature vectors are then used for recognition of facial expression by comparison using a normalized dot product similarity metric.

Experiments: Physical Model Method By computing the muscle activation feature vectors for each of 6 subjects making about 5 expressions each, we can assess the recognition accuracy of this physical-model based expression recognition method.

Figure 15 shows the peak muscle actuations for 6 people making different expressions. The muscle template used for recognition is also shown for comparison. The dot products of the feature vectors with

[2]It may be that a better way to account for temporal aspects of facial motion would be to use a phase-portrait approach [26]. Although we are investigating such methods, at present our results suggest little need for incorporating such detail.

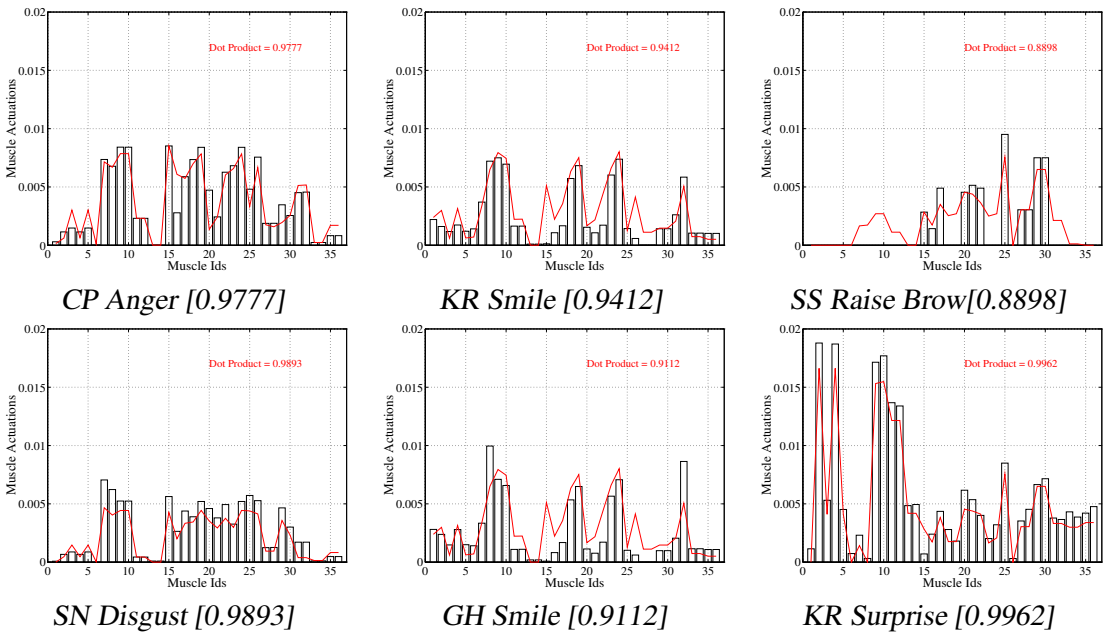


Figure 15: *Peak muscle actuations for several different expressions by different people. The dotted line shows the muscle template used for recognition. The normalized dot product of the feature vector with the muscle template is shown below each figure.*

the corresponding expression's muscle template are shown below each figure.

This figure is important in that it illustrates the repeatability and accuracy with which we can estimate muscle actuations. The largest differences in peak muscle actuations are due to facial asymmetry and intensity of the actuation. The intensity difference is especially apparent in the case of the surprise expression where some people open their mouth less then others. Our analysis does not enforce any symmetry constraints and none of our data, including the muscle templates shown in Figure 14, portray exactly symmetric expressions.

Table 1 shows the results of dot products between peak muscle actuations of five randomly chosen expressions with each expression's muscle template. It can be seen that for the five instances shown in this table, each expression is correctly identified.

We have used this recognition method with all of our test data, consisting of 8 different people making the 52 expressions (including: *smile*, *surprise*, *raise eye brow*, *anger*, and *disgust*) Since some people did not make all expressions, we have 10 samples each for surprise, anger, disgust, and raise eyebrow expressions, and 12 for the smile expression. Table 2 shows the average and the standard deviation of all the expressions compared to each expression's muscle template. Of particular note is the excellent repeatability and consistency of the muscle activation estimates.

Table 3 shows the results of the overall recognition results in the form of a *classification matrix*. In our tests there was only one recognition failure, for the expression of anger. Our overall accuracy was 98.0%.

	Smile	Surprise	Anger	Disgust	Raise Brow
Smile	**0.91**	0.36	0.91	0.75	0.17
Surprise	0.32	**0.99**	0.34	0.28	0.20
Anger	0.62	0.28	**0.99**	0.47	0.81
Disgust	0.32	0.22	0.66	**0.88**	0.43
Raise Brow	0.75	0.27	0.84	0.24	**0.98**

Table 1: *Some examples of recognition of facial expressions, using peak muscles actuations. A score of 1.0 indicates complete similarity.*

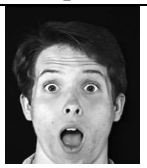
Expressions	Smile	Surprise	Anger	Disgust	Raise Brow
Template					
Smile	**0.97 ± 0.03**	0.63 ± 0.04	0.95 ± 0.01	0.86 ± 0.04	0.59 ± 0.16
Surprise	0.58 ± 0.03	**0.99 ± 0.01**	0.59 ± 0.04	0.57 ± 0.05	0.56 ± 0.09
Anger	0.90 ± 0.05	0.55 ± 0.05	**0.97 ± 0.02**	0.91 ± 0.01	0.65 ± 0.14
Disgust	0.82 ± 0.06	0.57 ± 0.05	0.92 ± 0.03	**0.95 ± 0.03**	0.78 ± 0.10
Raise Brow	0.58 ± 0.05	0.57 ± 0.07	0.70 ± 0.05	0.78 ± 0.06	**0.96 ± 0.04**

Table 2: *Mean and Standard Deviations of Similarity scores of all expressions in the database. Similarity metric is normalized dot products*

4.2 Spatio-temporal Motion Energy Templates for Recognition

The previous section used estimated peak muscle actuations as a feature to detect similarity/dissimilarity of facial expressions. Now we consider a second, much simpler representation: the templates of facial motion energy.

Figure 16 shows the pattern of motion generated by averaging two randomly chosen subjects per expression from our facial expression image sequence database. Notice that each of these motion templates is unique and therefore can serve as an sufficient feature for categorization of facial expression. Note also that these motion-energy templates are sufficiently smooth that they can be subsampled at one-tenth the raw image resolution, greatly reducing the computational load.

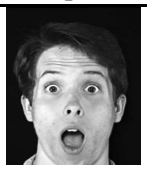
Expressions	Smile	Surprise	Anger	Disgust	Raise Brow
Template					
Smile	**12**	0	1	0	0
Surprise	0	**10**	0	0	0
Anger	0	0	**9**	0	0
Disgust	0	0	0	**10**	0
Raise Brow	0	0	0	0	**10**
Success	100%	100%	90%	100%	100%

Table 3: *Results of Facial Expression Recognition using peak-muscle actuations. This result is on based on 12 image sequences of smile, 10 image sequences of surprise, anger, disgust, and raise eyebrow. Success rate for each expression is shown in the bottom row. The overall recognition rate is 98.0%.*

We use the Euclidean norm of the difference of between the motion energy template and standard 2-D optical flow motion energy as a metric for measuring the similarity/dissimilarity of expressions. Note that metric works oppositely from the dot-product metric: the lower the value of this metric, more similar the expressions.

Experiments: Spatio-temporal motion energy templates Using the average of two people making an expression, we generate motion-energy template images for each of the five expressions. Using these templates (shown in Figure 16), we conducted recognition tests for our whole database of 52 image sequences. Figure 17 shows six examples of the motion energy images generated by different people. The similarity scores of these to the corresponding expression templates are shown below each figure.

Table 4 shows the results of this recognition test for five different expressions by different subjects. The scores show that all the five expressions were correctly identified. Conducting this analysis for the whole database of 52 expressions, we can generate a table which shows the mean and variance of the similarity scores across the whole database. These scores are shown in Table 5. The classification results of this methods over the whole database, displayed as a *confusion/classification matrix*, are shown in Table 6. This table shows that again we have just one incorrect classification of the anger expression. The overall recognition rate with this method is also 98.0%.

5 Discussion and Conclusions

In this paper we have presented two methods for representation of facial motion. Unlike previous efforts at facial expression characterization, coding, interpretation, or recognition that have been based on the Facial Action Coding System (FACS), we have developed new, more accurate representations of facial motion and then used these new representations for the coding, and recognition/identification task.

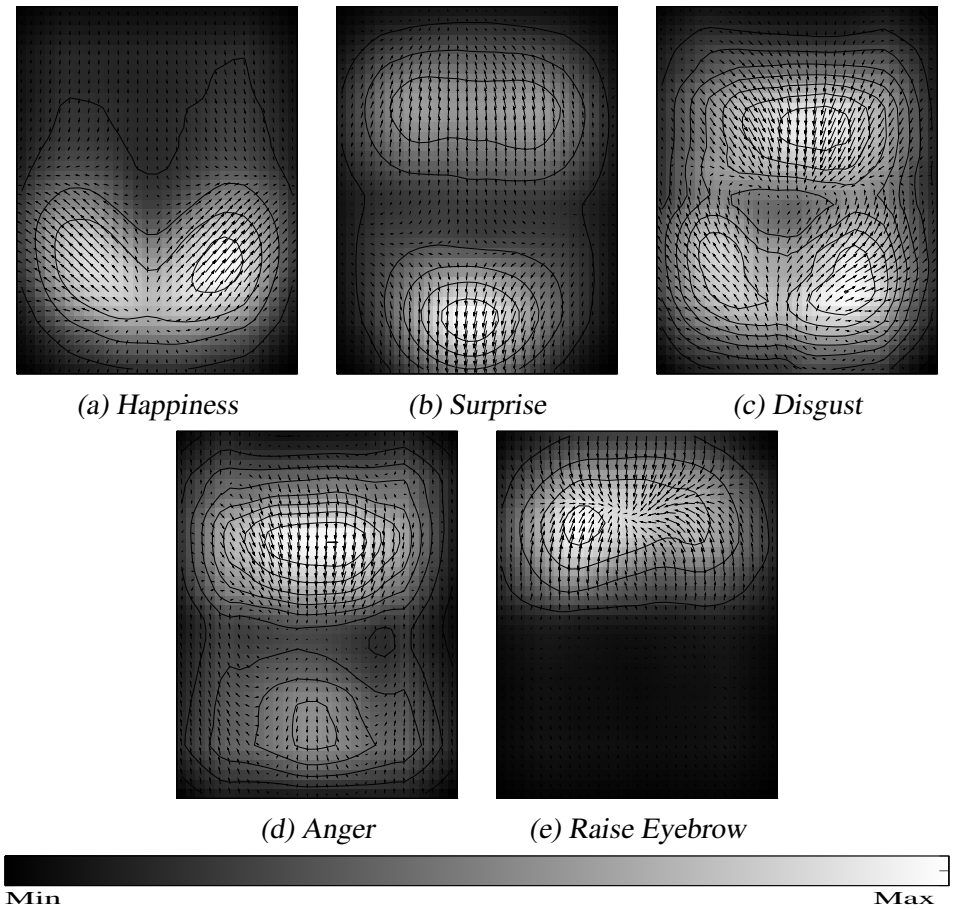


Figure 16: *Spatio-temporal motion-energy templates for the five expressions, averaged using the data from two people.*

We accomplished this by analyzing image sequences of facial expressions and then probabilistically characterizing the facial muscle activation associated with each expression. This is achieved using a detailed physics-based dynamic model of the skin and muscles coupled with optimal estimates of optical flow in a feedback controlled framework. This analysis produces a muscle-based representations of facial motion, which is then used to recognize facial expressions in two different ways.

The first recognition method we described uses the physics-based model directly, by recognizing expressions through comparison of estimated muscle activations. This method yields a recognition rate of 98% over our database of 52 sequences.

The second method uses the physics-based model to generate spatio-temporal motion-energy templates of the whole face for each different expression. These simple, biologically-plausible motion energy "templates" are then used for recognition. This method also yields a recognition rate of 98%. This level of accuracy at expression recognition is substantially better than has been previously achieved.

We have also used this representation in real-time tracking and synthesis of facial expressions [11]. We

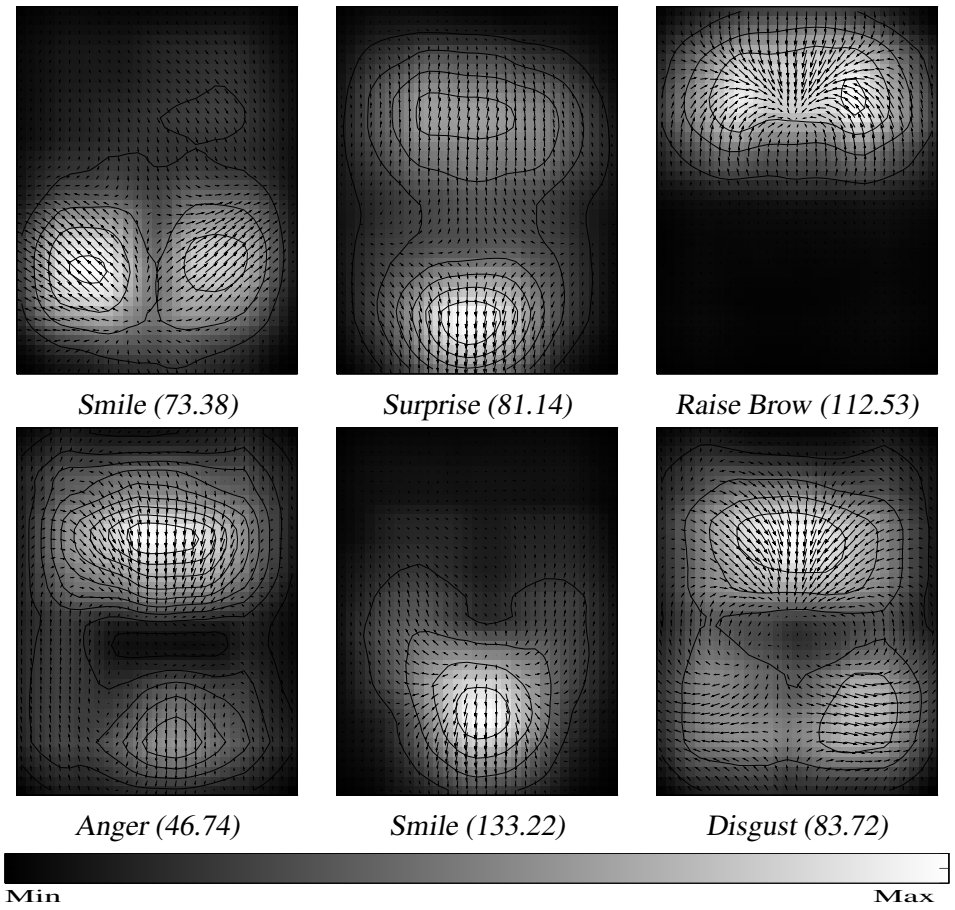


Figure 17: *Motion templates for four expressions from 4 people. Their similarity scores are also shown.*

are at the moment working on increasing the size of our database to also include other expressions and speech motions. We are also looking into model-based coding applications, biomedical applications and controlled experiments with psychologists.

References

[1] J. N. Bassili. Facial motion in the perception of faces and of emotional expression. *Journal of Experimental Psyschology*, 4:373–379, 1978.

[2] J. N. Bassili. Emotion recognition: The role of facial motion and the relative importance of upper and lower areas of the face. *Journal of Personality and Social Psychology*, 37:2049–2059, 1979.

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

Expressions	Smile	Surprise	Anger	Disgust	Raise Brow
Template					
Smile	**73.4**	255.0	230.3	209.4	294.6
Surprise	233.1	**81.1**	143.7	141.4	243.4
Anger	213.8	187.0	**46.7**	95.2	152.3
Disgust	154.0	178.3	126.4	**83.7**	227.3
Raise Brow	288.9	322.2	147.7	240.1	**46.8**

Table 4: *Example scores for recognition of facial expressions using spatio-temporal templates. Low Scores show more similarity to the template.*

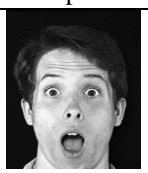
Expressions	Smile	Surprise	Anger	Disgust	Raise Brow
Template					
Smile	**94.1±34.7**	266.2 ± 52.3	234.5 ± 62.7	153.7 ± 59.7	306.6 ± 15.3
Surprise	230.9 ± 8.7	**123.6±70.7**	160.5 ± 38.3	173.5 ± 14.2	233.4 ± 14.1
Anger	225.7 ± 16.5	199.2 ± 76.0	**98.3±46.3**	160.1 ± 29.1	147.0 ± 15.5
Disgust	149.0 ± 22.7	198.1 ± 54.0	140.3 ± 43.7	**99.3±23.4**	224.3 ± 16.2
Raise Brow	339.9 ± 32.9	321.6 ± 96.4	208.9 ± 33.0	293.2 ± 26.8	**106.8±27.0**

Table 5: *Mean ± Standard Deviation of scores for recognition of facial expressions using the spatio-temporal templates over the whole database. Low Scores show more similarity to the template.*

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

[5] V. Bruce. *Recognising Faces*. Lawrence Erlbaum Associates, 1988.

[6] J. S. Bruner and R. Taguiri. The perception of people. In *Handbook of Social Psyschology*. Addison-Wesley, 1954.

[7] C. Darwin. *The expression of the emotions in man and animals*. University of Chicago Press, 1965. (Original work published in 1872).

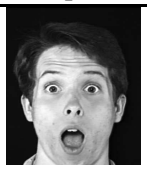
Expressions	Smile	Surprise	Anger	Disgust	Raise Brow
Template					
Smile	**12**	0	0	0	0
Surprise	0	**10**	0	0	0
Anger	0	0	**9**	0	0
Disgust	0	0	1	**10**	0
Raise Brow	0	0	0	0	**8**
Success	100%	100%	90%	100%	100%

Table 6: *Results of Facial Expression Recognition using spatio-temporal motion energy templates. This result is on based on 12 image sequences of smile, 10 image sequences of surprise, anger, disgust, and raise eyebrow. Success rate for each expression is shown in the bottom row. The overall recognition rate is 98.0%.*

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

[9] 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.

[10] I. Essa. *Analysis, Interpretation, and Synthesis of Facial Expressions*. PhD thesis, Massachusetts Institute of Technology, MIT Media Laboratory, Cambridge, MA 02139, USA, 1994.

[11] 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.

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

[13] B. Friedland. *Control System Design: An Introduction to State-Space Methods*. McGraw-Hill, 1986.

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

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

[16] K. Mase and A. Pentland. Lipreading by optical flow. *Systems and Computers*, 22(6):67–76, 1991.

[17] D. Metaxas and D. Terzopoulos. Shape and nonrigid motion estimation through physics-based synthesis. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 15(6):581–591, 1993.

[18] M. Minsky. *The Society of Mind*. A Touchstone Book, Simon and Schuster Inc., 1985.

[19] 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.

[20] 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.

[21] 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.

[22] A. Pentland and S. Sclaroff. Closed form solutions for physically based shape modeling and recovery. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 13(7):715–729, July 1991.

[23] 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.

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

[25] M. Rosenblum, Y. Yacoob, and L. Davis. Human emotion recognition from motion using a radial basis function network architecture. In *The Workshop on Motion of Nonrigid and Articulated Objects*, pages 43–49. IEEE Computer Society, 1994.

[26] E. Shavit and A. Jepson. Motion understanding using phase portraits. In *Looking at People Workshop*. IJCAI, 1993.

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

[28] 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.

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

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

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

[32] Y. Yacoob and L. Davis. Computing spatio-temporal representations of human faces. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 70–75. IEEE Computer Society, 1994.