Georgia Institute of Technology, Graphics, Visualization and Usability Center, Technical Report # GIT-GVU-98-23, August 1998.

An abridged version of this paper appears in IEEE Transactions on Pattern Analysis and Machine Intelligence,Volume 19 (7), July, 1997.

Coding, Analysis, Interpretation, and Recognition of Facial Expressions

Irfan A. Essa

College of Computing, GVU Center,
Georgia Institute of Technology,
Atlanta, GA 30332-0280, USA.
irfan@cc.gatech.edu.

Abstract

We describe a computer vision system for observing facial motion by using an optimal estimation *optical flow method coupled with geometric, physical and motion-based dynamic models 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+.

Finally, we show how this method can be used for coding, analysis, interpretation, and recognition of facial expressions.

Keywords: Facial Expression Analysis, Expression Recognition, Face Processing, Emotion Recognition, Facial Analysis, Motion Analysis, Perception of Action, Vision-based HCI.

1. Introduction

Faces are much more than keys to individual identity, they play a major role in communication and interaction that makes machine understanding, perception and modeling 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 [10, 23] 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? [6]. Although a large body of work dealing with human perception of facial motions exists, there have been few attempt to develop objective methods for quantifying facial movements.

Perhaps the most important work in this area is that of Ekman and Friesen [9], who 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.

However a well recognized limitation of this method is the lack of temporal and detailed spatial information (both at local and global scales) [10, 23]. Additionally, the heuristic "dictionary" of facial actions originally developed for FACS-based coding of emotion, after initial experimentation, has proven quite difficult to adapt for machine recognition of facial expression.

To improve this situation we would like to *objectively* quantify facial movements 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 dynamic model of motion, may it be physics-based model of both skin and muscle, geometric representation of a face or a motion specific model.

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 coarticulations, and relates facial motions to facial expressions. We will further demonstrate that the parameters extracted using this method provide improved accuracy for analysis, interpretation, coding and recognition of facial expression.

1.1 Background

Representations of Facial Motion: Ekman and Friesen [9] 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 result in a large set of possible facial expressions. For example smile 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. Despite its limitations this method is the most widely used method for measuring human facial motion for both human and machine perception.

Tracking facial motion: There have been several attempts to track facial expressions over time. Mase and Pentland [20] 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 [29] developed a much more sophisticated method that tracked linear facial features to estimate corresponding parameters of a three dimensional wire-frame 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.

Haibo Li, Pertti Roivainen and Robert Forchheimer [18] describe an approach in which a control feedback loop between what is being visualized and what is being analyzed is used for a facial image coding system. Their work is the most similar to ours, but both our goals and implementation are different. The main limitation of their work is 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 severe; it means we cannot observe the "true" patterns of dynamic model changes (*i.e.,* muscle actuations) because the method assumes the FACS model as the underlying representation. We are also interested in developing a representation that is not dependent on FACS and is suitable not just for tracking, but recognition and analysis.

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 several researchers [19, 20, 33, 4] for their recognition systems. Yacoob and Davis, who extend the work of Mase, detect motion (only in eight directions) in six predefined and hand initialized rectangular regions on a face and then use simplifications of the FACS rules for the six universal expressions for recognition. The motion in these rectangular regions, from the last several frames, is correlated to the FACS rules for recognition. Black and Yacoob extend this method, using local parameterized models of image motion to deal with large-scale head motions. These methods show about 90% accuracy at recognizing expressions in a database of 105 expressions [4, 33]. Mase [19] on a smaller set of data (30 test cases) 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. Several researchers [1, 2, 6, 7, 8, 17] have claimed that the timing of expressions, something that is completely missing from FACS, is a critical parameter in recognizing emotions. This issue was also addressed in the NSF workshops and reports on facial expressions [10, 23]. 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.

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 have found that Simoncelli's [28] method for optical flow computation, which uses a multi-scale, coarse-to-fine, Kalman filtering-based algorithm, 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 $\mathbf{\Lambda_v}$ between different frames for determining confidence measures and for error corrections in observations for the dynamic model (see Section 2.3 and Figure 3 [observation loop (a)]).

2.2 Facial Modeling

A priori information about facial structure is required for our framework. We use a face model which is an elaboration of the facial mesh developed by Platt and Badler [27]. We extend this into a topologically invariant physics-based model by adding anatomically-based muscles to it [11].

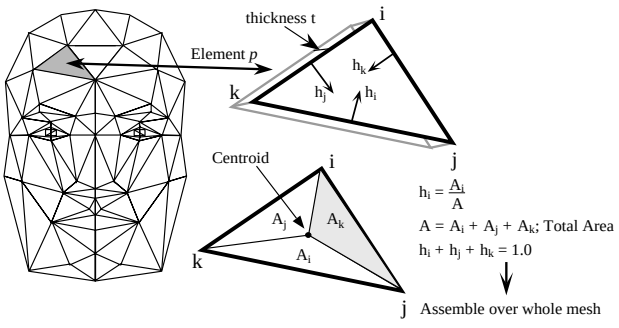

Figure 1. *Using the geometric mesh to determine the continuum mechanics parameters of the skin using Finite Element Methods.*

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.

By modeling the elastic nature of facial skin and the anatomical nature of facial muscles we develop a dynamic muscle-based model of the face, including FACS-like control parameters (see [11, 32] for implementation details). 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 [3, 15, 21].

By defining each of the triangles on the polygonal mesh as an *isoparametric triangular shell element*, (shown in Figure 1), we can calculate the mass, stiffness and damping matrices for each element (using $dV = t_{el} dA$), where t_{el} is thickness, given the material properties of skin (acquired from [26, 30]). Then by the assemblage process of the direct stiffness method [3, 15] 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_{el}, 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 [26] and Waters [32] 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. Initially we started our estimation process by manually translating, rotating and deforming our 3-D facial model to fit a face in an image. To automate this process we are now using the View-based and Modular Eigenspace methods of Pentland and Moghaddam [22, 24].

Using this method we can automatically extract the positions of the eyes, nose and lips in an image as shown in Figure 2(b). These feature positions are used to warp the face image to match the canonical face mesh (Figure 2(c) and (d)). This allows us to extract the additional "canonical feature points" on the image that correspond to the fixed (non-rigid) nodes on our mesh (Figure 2(f)). After the initial registering of the model to the image the coarse-to-fine flow computation methods presented by Simoncelli [28] and Wang [31] 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. This limitation can be addressed by using methods that attempt to track and stabilize head movements (*e.g.,* [12, 4]).

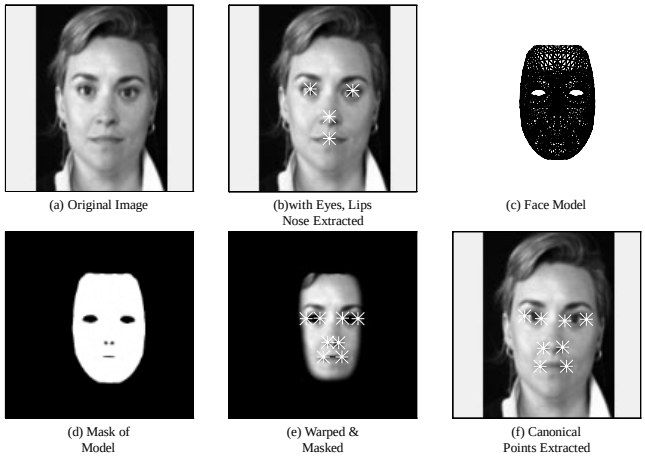

Figure 2. *Initialization on a face image using Modular Eigenfeatures method with a canonical model of a face.*

Images to face model

Simoncelli's [28] 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, $\mathcal{M}$, we would like to compute velocities for the vertices of the face model $\mathbf{v}_g$. Then, using the physically-based modeling techniques and the relevant geometric and physical models, described earlier, 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 $\mathcal{T}$), and rotations (matrix $\mathcal{R}$). The Galerkin polynomial interpolation function $\mathbf{H}$ and the strain-displacement function $\mathcal{B}$, 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 [15, 25, 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, $\mathcal{S}(u, v)$, 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:

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

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.

Estimation and Control

Driving a physical system with the inputs from noisy motion estimates can result in divergence or a chaotic physical response.

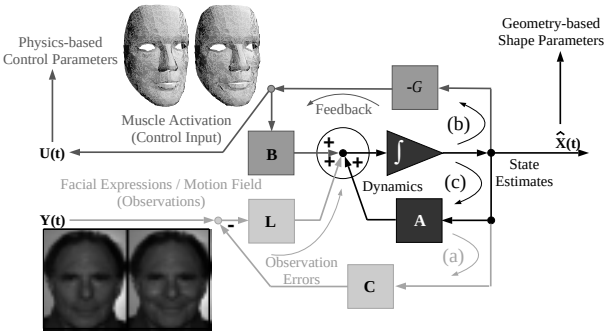

Figure 3. *Block diagram of the control-theoretic approach. Showing the estimation and correction loop (a), the dynamics loop (b), and the feedback loop (c).*

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 [18]. Figure 3 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 [5, 16]. 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:

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

where $\hat{\mathbf{X}}$ is the linear least squares estimate of $\mathbf{X}$ based on $Y(\tau)$ for $\tau < t$ and $\mathbf{\Lambda}_e$ 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 $\mathbf{\Lambda}_e$:

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

We solve for $\mathbf{\Lambda}_e$ in Equation (3) assuming a steady-state system (*i.e.,* $\frac{d}{dt}\mathbf{\Lambda}_e = 0$).

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 3 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 ($\mathbf{\Lambda}_\mathbf{v}(t)$) with respect to the observations. We can simply use this distribution in our dynamic observations relationships. Hence we obtain:

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

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: $\mathbf{U} = -\mathcal{G}\mathbf{X}$, where $\mathcal{G}$ is the *control feedback gain matrix*. We assume the instance of control under study falls into the category of an *optimal regulator* (as we need some optimality criteria for an optimal solution [16]). 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* [16]. 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 $\mathcal{G} = \mathbf{R}^{-1}\mathbf{B}^T\mathbf{P}_c$. This control loop is also shown in the block diagram in Figure 3 (upper loop (c)).

2.4 Motion Templates from the Facial Model

So far we have discussed how we can extract the muscle actuations of an observed expression. These methods have relied on detailed geometric and/or physics-based description of facial structure. 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. In other words, as much as the dynamics of motion is implicit into our analysis, it does not explicitly require a geometric and/or physical model of the structure. The detailed models are there so that we can *back-project* the facial motion onto these models and use these models to extract a representation in the state-space of these models. We could just use the motion and velocity measurements for analysis, interpretation and recognition without using the geometric/physical models. This is possible by using 2-D motion energy templates that encode just the motion. This encoded motion in 2-D is then used as representation for facial action.

The system shown in Figure 3 employs optimal estimation, within an optimal control and feedback framework. It maps 2-D motion observations from images onto a 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 9 and Figure 10 show the corrected flow for the expressions of raise eyebrow and smile, and also show the corrected flow after it has been applied to a dynamic face model. Further corrections are possible by using deformations of the facial skin (*i.e.,* the physics-based model) as constraints in state-space that only measures image motion.

By using this methodology without the detailed 3-D geometric and physical models and *back-projecting* the facial motion estimates into the image we can remove the complexity of physics-based modeling from our representation of facial motion. Then learning the "ideal" 2-D motion views (*e.g.,* motion energy) for each expression we can characterize spatio-temporal templates for those expressions. Figure 4 (e) and (f) shows examples of

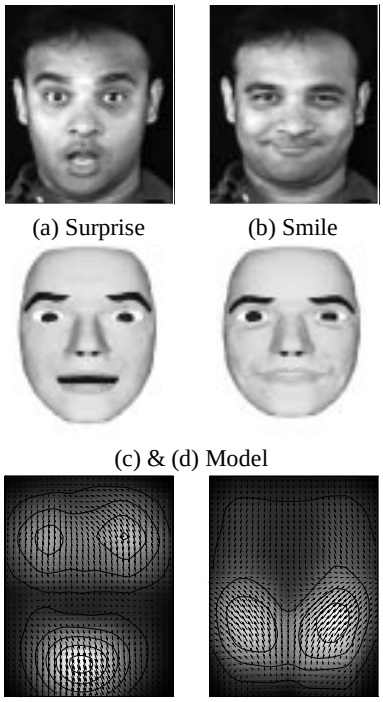

Figure 4. *Determining of expressions from video sequences. (a) and (b) show expressions of smile and surprise, (c) and (d) show a 3-D model with surprise and smile expressions, and (e) and (f) show the spatio-temporal motion energy representation of facial motion for these expressions.*

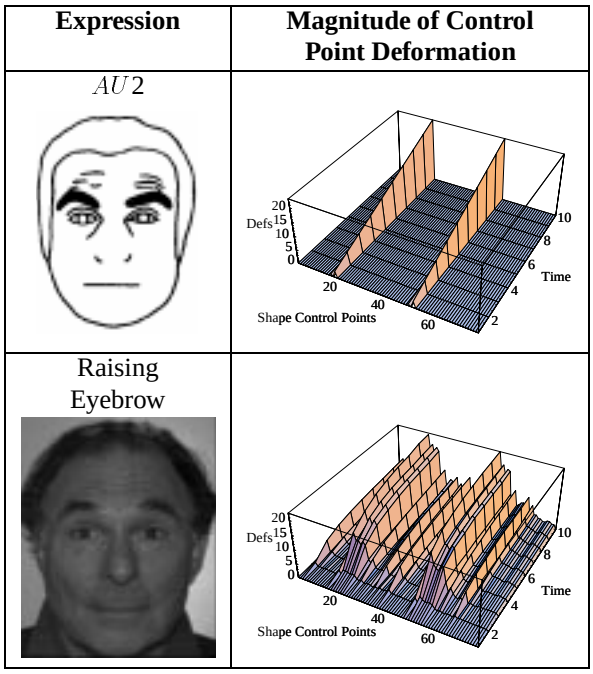

Figure 5. *FACS/Candide deformation vs. Observed deformation for the Raising Eyebrow expression. Surface plots (top) show deformation over time for FACS actions $AU2$, and (bottom) for an actual video sequence of raising eyebrows.*

this representation of facial motion energy. It is this representation of facial motion that we use for generating spatio-temporal "templates" for coding, interpretation and recognition of facial expressions.

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) have 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 [32], 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.

Evaluation is an important part of our work as we do need to experiment extensively on real data to measure the validity of our new representation. For this purpose we have developed a video database of people making expressions;the results presented here are based on 52 video sequences of 8 users making 6 different expressions. These expressions were all acquired at 30 frames per second at full NTSC video resolution.

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 measuring facial motion and not human emotion. 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 also be briefly ad-

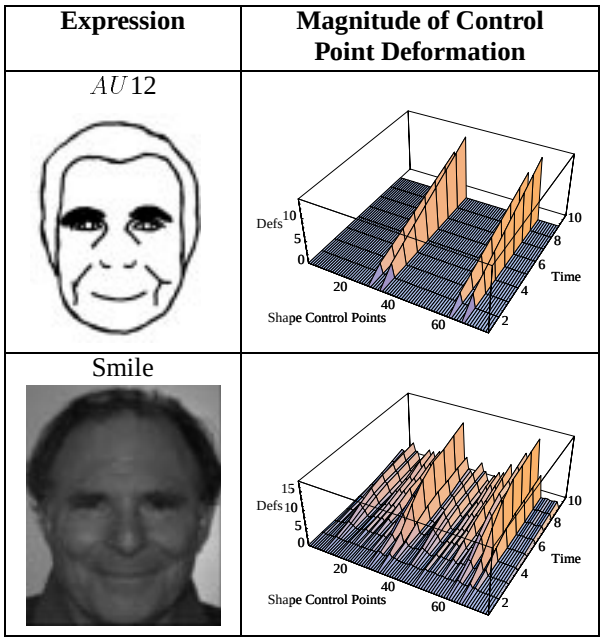

Figure 6. *FACS/Candide deformation vs. Observed deformation for the Happiness expression. Surface plots (top) show deformation over time for FACS action $AU12$, and (bottom) for an actual video sequence of happiness.*

dressed.

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 5 and Figure 6.

The top row of Figure 5 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 5 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 smile expression are shown in Figure 6. 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.

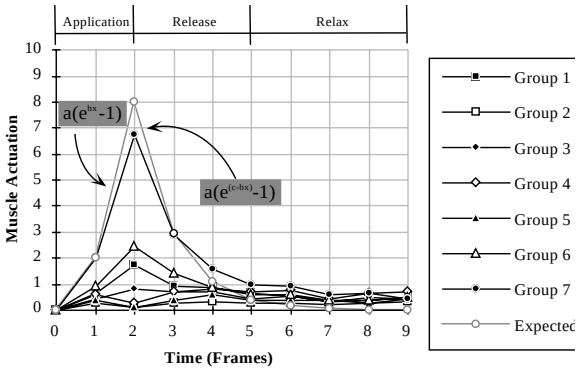

Figure 7. *Actuations over time of the seven main muscle groups for the expressions of raising brow. The plots shows actuations over time for the seven muscle groups and the expected profile of application, release and relax phases of muscle activation.*

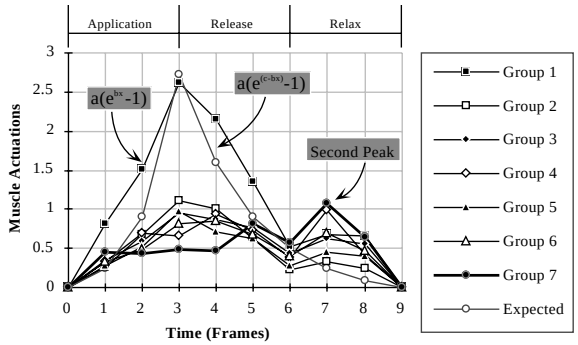

Figure 8. *Actuations over time of the seven main muscle groups for the expressions of smiling – lip motion. The plots shows actuations over time for the seven muscle groups and the expected profile of application, release and relax phases of muscle activation.*

3.2 Temporal Patterning

Another important observation about facial motion that is apparent in Figure 5 and Figure 6 is that the facial motion is far from linear in time. This observation becomes much more important when facial motion is 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 5 and Figure 6, that show the development of FACS expressions can only be represented by a muscle actuation that has a step-function profile. Figure 7 and Figure 8 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 8 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 7 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 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.

Our initial experimentation on automatic characterization of facial expression is based on 52 image sequences of 8 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 study. Complete detail of our work on expression recognition using the representations discussed here appears elsewhere [14]. Using two different methods; one based on our detailed physical model and the other on our 2-D spatio-temporal motion energy templates, both showed recognition accuracy rates of 98%.

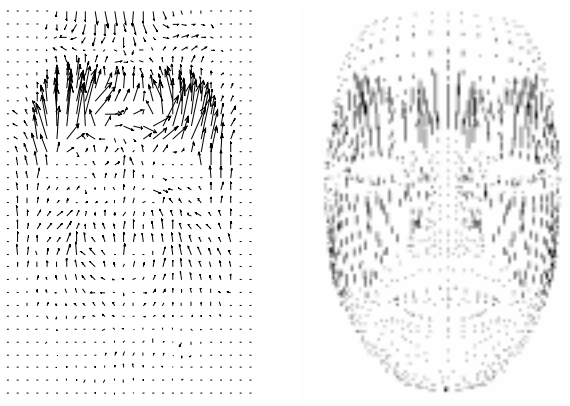

Figure 9. *Left figure shows a motion field for the expression of raise eye brow expression from optical flow computation and the right figures shows the motion field after it has been mapped to a dynamic face model using the control-theoretic approach of Figure 3.*

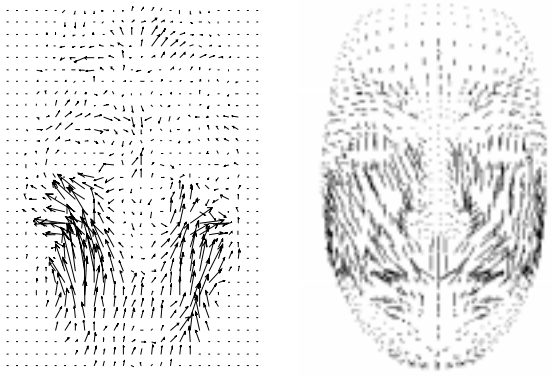

Figure 10. *Left figure shows a motion field for the expression of smile expression from optical flow computation and the right figures shows the motion field after it has been mapped to a dynamic face model using the control-theoretic approach of Figure 3.*

4. Discussion and Conclusions

We have developed a mathematical formulation and implemented a computer vision system capable of detailed analysis of facial expressions within an active and dynamic framework. The purpose of this system to to analyze real facial motion in order to derive an improved model (FACS+) of the spatial and temporal patterns exhibited by the human face.

This system analyzes facial expressions by observing expressive articulations of a subject's face in video sequences. The visual observation (sensing) is achieved by using an *optimal optical flow* method. This motion is then coupled to a physical model describing the skin and muscle structure, and the muscle control variables estimated.

By observing the control parameters over a wide range of facial motions, we can then extract a minimal parametric representation of facial control. We can also extract a minimal parametric representation of facial patterning, a representation useful for static analysis of facial expression.

We have used this representation in real-time tracking and synthesis of facial expressions [13] and have experimented with expression recognition. Currently our expression recognition accuracy is 98% on a database of 52 sequences. using either our muscle models or 2-D motion energy models for classification [14].

We are working on expanding our database to cover many other expressions and also expressions with speech. 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 and procuring funding to undertake controlled experiments in the area with

more emphasis on evaluation and validity.

Acknowledgments

We would like to thank Eero Simoncelli, John Wang Trevor Darrell, Paul Ekman, Steve Pieper, and Keith Waters, Steve Platt, Norm Badler and Nancy Etcoff for their help with various parts of this project.

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.

[4] 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.

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

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

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

[8] P. Ekman. The argument and evidence about universals in facial expressions of emotion. In H. Wagner and A. Manstead, editors, *Handbook of Social Psychophysiology*. Lawrence Erlbaum, 1989.

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

[10] 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.

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

[12] I. Essa, S. Basu, T. Darrell, and A. Pentland. Modeling, tracking and interactive animation of faces and heads using input from video. In *Proceedings of Computer Animation Conference 1996*, pages 68–79. IEEE Computer Society Press, June 1996.

[13] 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.

[14] I. Essa and A. Pentland. Facial expression recognition using a dynamic model and motion energy. In *Proceedings of the International Conference on Computer Vision*, pages 360–367. IEEE Computer Society, Cambridge, MA, 1995.

[15] I. A. Essa, S. Sclaroff, and A. Pentland. A unified approach for physical and geometric modeling for graphics and animation. *Computer Graphics Forum, The International Journal of the Eurographics Association*, 2(3), 1992.

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

[17] C. E. Izard. Facial expressions and the regulation of emotions. *Journal of Personality and Socail Psychology*, 58(3):487–498, 1990.

[18] 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.

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

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

[21] 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.

[22] 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.

[23] 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.

[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.

[25] 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.

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

[29] 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.

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

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

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

[33] 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.