International Journal of Computer Vision 25(1), 23–48 (1997)

Recognizing Facial Expressions in Image Sequences Using Local Parameterized Models of Image Motion

MICHAEL J. BLACK
Xerox Palo Alto Research Center, 3333 Coyote Hill Road, Palo Alto, CA 94304
black@parc.xerox.com

YASER YACOOB
Computer Vision Laboratory, University of Maryland, College Park, MD 20742
yaser@cs.umd.edu

Received March 27, 1995; Revised June 4, 1996; Accepted August 9, 1996

Abstract. This paper explores the use of local parametrized models of image motion for recovering and recognizing the non-rigid and articulated motion of human faces. Parametric flow models (for example affine) are popular for estimating motion in rigid scenes. We observe that within local regions in space and time, such models not only accurately model non-rigid facial motions but also provide a concise description of the motion in terms of a small number of parameters. These parameters are intuitively related to the motion of facial features during facial expressions and we show how expressions such as anger, happiness, surprise, fear, disgust, and sadness can be recognized from the local parametric motions in the presence of significant head motion. The motion tracking and expression recognition approach performed with high accuracy in extensive laboratory experiments involving 40 subjects as well as in television and movie sequences.

Keywords: facial expression recognition, optical flow, parametric models of image motion, robust estimation, non-rigid motion, image sequences

1. Introduction

The recognition of facial expressions in image sequences with significant head motion is a challenging problem with many applications for human-computer interaction. Yet, while the coincidence of head and facial feature motion is prevalent in human behavior, it has so far attracted only little attention as a motion estimation problem. Previous work has typically focused on one part of the problem or the other: either rigid head tracking (Azarbayejani et al., 1993a, 1993b) with no facial expressions or expression recognition with either no motion at all (Yuille and Hallinan, 1992) or a roughly stationary head with a changing expression (Terzopoulos and Waters, 1993; Yacoob and Davis, 1994). Here we propose a simple model of rigid and non-rigid facial motion using a collection of local parametric models. The image motions of the face, mouth, eyebrows, and eyes are modeled using image flow models with only a handful of parameters. The motions of these regions are estimated over an image sequence using a robust regression scheme (Black and Anandan, 1996) which makes the recovered motion parameters stable under adverse conditions such as motion blur, saturation, loss of focus, etc. These recovered parameters correspond simply and intuitively to various facial expressions. We illustrate how the motion parameters can be used to recognize facial expressions even in situations where the motion of the head is large.

Models used in recognizing facial expressions vary in the amount of geometric information about head shape and motion they contain. At one extreme are approaches which employ physically-based models of heads including skin and musculature (Essa and Pentland, 1994; Terzopoulos and Waters, 1993). A slightly weaker model uses deformable templates to represent feature shapes in the image (Yuille and Hallinan, 1992). At the other extreme is the work of Yacoob and Davis (1994) in which they recognize facial expressions using statistical properties of the optical flow with only very weak models of facial shape. In this paper we explore a middle ground between the template-based approaches and the optical flow-based approaches in which we represent rigid and deformable facial motions using piecewise parametric models of image motion. These models provide greater abstraction and robustness than the purely flow-based methods yet are weaker than models which incorporate detailed information about shape. While parametric flow models (for example affine) are popular for estimating motion in rigid scenes (Bergen et al., 1992; Black and Anandan, 1996; Black and Jepson, 1994), their application to non-rigid motion is unconventional. However, within local regions in space and time, such models not only accurately model non-rigid facial motions but also provide a concise description of the motion in terms of a small number of parameters.

To model the rigid motion of a face in the image we make the simple assumption that the majority of the face can be modeled by a plane. More complex models can be employed (for example an ellipsoid) but the planar model is particularly simple as the image motion of a plane can be described by eight parameters. A face is neither planar nor strictly rigid but, when robust estimation techniques are employed, the simple planar model can be used in situations where there are outliers due to non-planarity or non-rigidity. This planar model is sufficient for recovering qualitative information about the motion of the head. More sophisticated models could be used if accurate information about the 3D motion of the head is required.

The image motions of the facial features (eyes, mouth and eyebrows) are modeled relative to the head motion using different parametric models. For the eyes a simple affine model is used. For the brows and mouth an affine model is augmented with an additional curvature parameter to account for the arching of the brows and curvature of the mouth during smiling. Additionally, the nose region can be tracked, if desired, using an affine model. In this paper we do not address the problem of initially locating the various facial features; this topic has been addressed in (Chow and Li, 1993; Yacoob and Davis, 1993; Yuille et al., 1989). Notice that while the motion of the entire face can be quite complicated, when it is broken down into parts, the motion of each part can be modeled very simply.

The approach is summarized as follows (and is illustrated in Fig. 1). Given the location of the face, eyes, brows, and mouth, estimate the rigid motion of the face region (excluding the deformable features) between two frames using a planar motion model. This estimation is performed using a robust statistical approach to cope with violations of the rigid plane assumption. The motion of the face is used to register the images via warping and then the relative motion of the feature regions is estimated in the coordinate frame of the face using exactly the same robust estimation procedure. The motion estimates of the face and features are used to predict their locations in the next frame and the

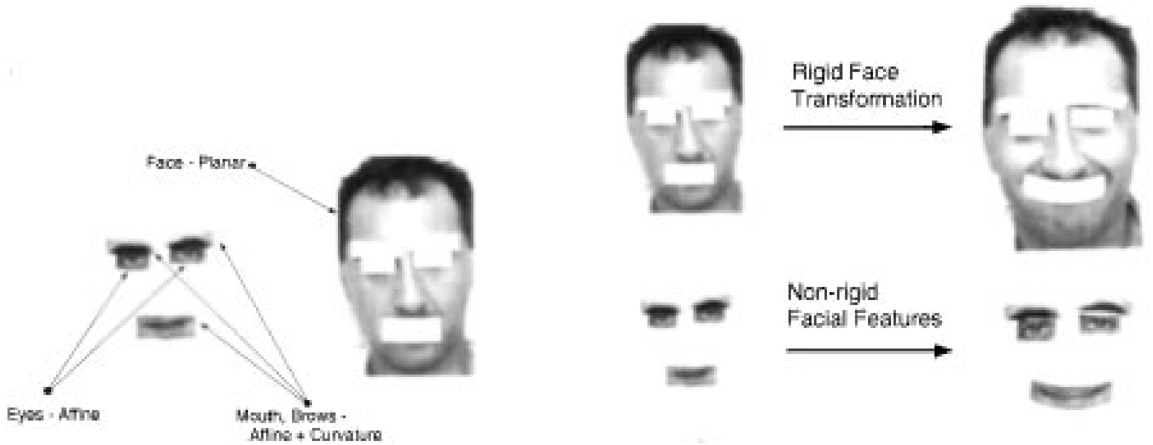

Face - Planar
Eyes - Affine
Mouth, Brows - Affine + Curvature
Rigid Face Transformation
Non-rigid Facial Features

Figure 1. Illustration showing the parametric motion models employed and an example of a face undergoing a looming motion while smiling.

process is repeated. The estimated motion parameters provide a simple abstraction of the underlying facial motions and can be used to classify the type of rigid head motion and the facial expression.

In the following section we review previous work on facial feature tracking and expression recognition. Section 3 presents the parametric motion models used for the various facial features and describes the robust estimation scheme used to recover the motion parameters. These parameters are related to our mid- and high-level representations of facial expressions in Section 4. Then Section 5 discusses the results of tracking experiments with various expressions and rigid head motions. The results of extensive testing of the recognition method are presented in Section 6 along with an analysis of the approach.

2. Previous Work

2.1. Human Facial Tracking

Head tracking involves tracking the motion of a rigid object performing rotations and translations while facial feature tracking involves tracking diverse non-rigid deformations that are limited by the anatomy of the head. There are two basic types of feature tracking: *feature boundary* and *feature region* tracking. Feature boundary tracking attempts to track and accurately delineate the shapes of the facial features—e.g., to track the contours of the lips and mouth opening (Kass et al., 1987; Terzopoulos and Waters, 1993; Blake and Isard, 1994). Feature region tracking, on the other hand, addresses the simpler problem of tracking a region encompassing the face feature, paying little if any attention to the detailed shape of the feature. In this paper we focus on the latter tracking approach, and show that it allows us to capture and describe several aspects of the rich repertoire of facial expressions.

Face features are subject to one or more of the following motions: rigid, articulated, and deformable motions. The rigid motion is due to the head's translation or rotation. The articulated motion includes the motion of the lower jaw during speech and several facial expressions (e.g., 'surprise' expression). Typical deformable motions are due to muscle contractions and expansions that accompany speech and facial expressions.

An approach for rigid head tracking and motion estimation by tracking points with high Hessian was proposed in (Azarbayejani et al., 1993b). Several such points are tracked over the head and the 3-D motion parameters of the head are recovered by solving an over-constrained set of motion equations. The approach does not deal with facial expressions.

Essa and Pentland (1994) proposed a 3-D model-based approach for tracking facial features. They assumed that a mesh was placed on the face and used the optical flow field to displace the mesh vertices and recover the location of points on the face during the facial deformation. Rigid head motion was not allowed as there was no way to factor the optical flow into separate head and feature motions.

In a related approach Essa et al. (1994) used a template-based strategy for recognizing facial expressions. Such an approach lacks explicit information about the motion of the features and may prove hard to generalize to situations with significant head motion. Head motion causes the appearance of the features to change thus requiring multiple templates to recognize the same expression under different viewing positions.

Li et al. (1993) proposed a model-based approach that assumes that a 3-D mesh has been placed on the face in the image, and that the depths of points on the face have been recovered. They proposed algorithms for recovering the rigid and non-rigid motions of the face from the sequence of images, and reapplied these motions to create an approximation to the initial sequence. Their model-based approach employed knowledge about the anatomy of the face to constrain the estimation of the non-rigid facial motion.

The facial feature tracking reported in Yacoob and Davis (1994) is based on analysis of the magnitudes of gradients of the intensity image and the optical flow fields of the image sequence. The changes in these values between consecutive images provided clues to the spatial change of each facial feature. This approach dealt well with articulated and deformable motions, but was able to accommodate only limited rigid motion.

The work reported in (Terzopoulos and Waters, 1993) assumes that eleven principal contours are initially located (in practice manually) on the face. These contours are tracked throughout the sequence by applying an image force field that is computed from the gradient of the intensity image. In addition to assuming a frontal view, it was assumed that the projection is orthographic and that some facial make-up is needed.

Two related approaches that allow face pose and expression estimation, and face tracking in the image plane were reported in (Beymer et al., 1993; Toelg and Pogio, 1994) respectively. The work reported in (Beymer et al., 1993) provides a method for learning of associations between images and head pose and facial expressions. This association can be used in both analysis and synthesis of face images although it does not explicitly capture facial deformation or pose changes. Toelg and Poggio (1994) proposed a hierarchical model-based representation to register faces. They perform coarse-to-fine estimation of an affine transformation for the face region between images. We have observed that a planar model of the face provides a better stabilization. Additionally, we separately model the relative motions of the facial features and use this motion information for expression recognition.

Mase (1991) approached facial expression recognition based on computing the motions of facial *muscles* from the optical flow rather than the motions of facial *features*. Four facial expressions were studied: surprise, anger, happiness, and disgust. Optical flow is computed within rectangles that include these muscle units, which in turn were related to facial expression. Feature vectors were defined over a 15-D space that is based on the means and variances of the optical flow and were used in expression classification.

Our objective is to develop a passive system that is able to track a human head and primary facial features in a dynamic environment and provide a rich description of the observed motions. The description we seek will allow separation between different classes of motion so that reasoning about facial expression and head gesture is feasible.

2.2. Expression Recognition

Research in psychology has indicated that at least six emotions are universally associated with distinct facial expressions (Ekman, 1992). Several other emotions, and many combinations of emotions, have been studied but remain unconfirmed as universally distinguishable. The six principal emotions are: happiness, sadness, surprise, fear, anger, and disgust.

Most psychological research on facial expressions has been conducted on "mug-shot" pictures that capture the subject's expression at its peak (Young and Ellis, 1989). These pictures allow one to detect the presence of static cues (such as wrinkles) as well as the position and shape of the facial features. Few studies have directly investigated the influence of the motion and deformation of facial features on the interpretation of facial expressions. Bassili (1979) suggested that motion in the image of a face would allow emotions to be identified even with minimal information about the spatial arrangement of features. The subjects of his experiments viewed image sequences in which only white dots on the dark surface of the face displaying the emotion are visible. The reported results indicate that facial expressions were more accurately recognized from dynamic images than from a single static image. Whereas all expressions were recognized at above chance levels in dynamic images, only happiness and sadness were recognized at above chance levels in static images.

Figure 2 summarizes the observations of Bassili (1979) on motion-based cues for facial expressions. Recall that the experiments of Bassili were intended to explore only the role of motion in facial expressions; therefore the face features, texture and complexion were unavailable to the subjects. As illustrated in Fig. 2, Bassili identified principal facial motions that provide powerful cues to the subjects to recognize facial expressions.

In developing our approach to expression recognition we rely on the psychology sources best represented for static imagery by the work of Ekman and Friesen (1975), and for motion images by the work of Bassili (1979), as well as on the recent mid and high level spatio-temporal representations proposed by Yacoob and Davis (1994).

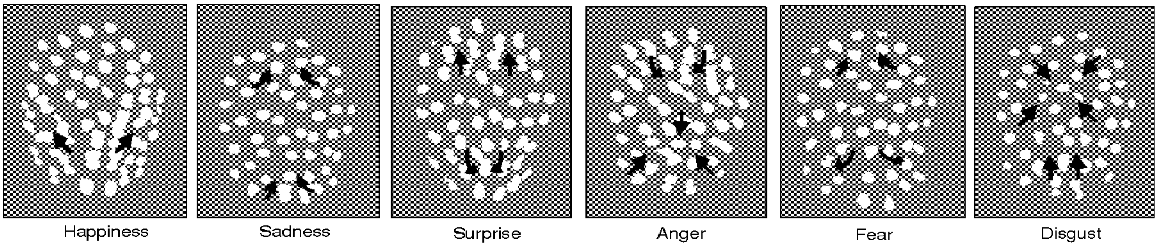


Figure 2. The cues for facial expression as suggested by Bassili.

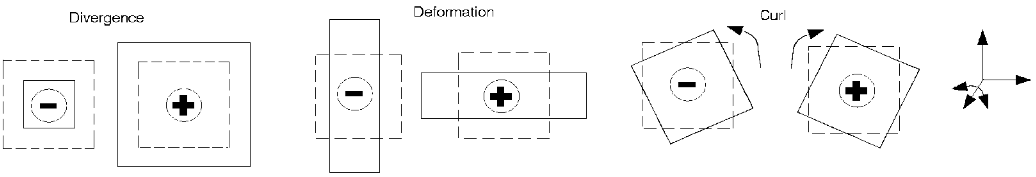


Figure 3. The figure illustrates the motion captured by the various parameters used to represent the motion of the regions. The solid lines indicate the deformed image region and the "−" and "+" indicate the sign of the quantity.

3. Estimating Facial Motions

This section introduces the parameterized optical flow models and the robust estimation procedure used to recover and track the rigid motion of the face and the relative motions of the facial features.

3.1. Motion Models

The estimation of image motion requires the integration of information over some neighborhood of the image under some assumptions about the variation of the motion. Parameterized models of image motion make explicit the assumptions about the motion and typically assume that the image flow can be represented by a low-order polynomial (Bergen et al., 1992). Within small image regions the following affine model of image motion is often sufficient (Koenderink and van Doorn, 1975)

$$u(x, y) = a_0 + a_1 x + a_2 y \qquad (1)$$

$$v(x, y) = a_3 + a_4 x + a_5 y \qquad (2)$$

where the a_i are constants, $\mathbf{u}(\mathbf{x}) = [u(x, y), v(x, y)]^T$ are the horizontal and vertical components of the flow at the image point $\mathbf{x} = (x, y)$, and the spatial positions $\mathbf{x}$ are defined with respect to some image point (typically the center of the region).

The parameters a_i have qualitative interpretations in terms of image motion. For example, a_0 and a_3 represent horizontal and vertical translation respectively. Additionally, we can express *divergence* (isotropic expansion), *curl* (rotation about the viewing direction), and *deformation* (squashing or stretching) as combinations of the a_i (Cipolla and Blake, 1992; Koenderink and van Doorn, 1975):

$$\text{divergence} = a_1 + a_5 = (u_x + v_y), \qquad (3)$$

$$\text{curl} = -a_2 + a_4 = -(u_y - v_x), \qquad (4)$$

$$\text{deformation} = a_1 - a_5 = (u_x - v_y) \qquad (5)$$

where the subscripts x and y indicate partial derivatives of the image velocity. Divergence, curl and the magnitude of the deformation have the convenient property of being invariant to rotations of the image coordinate frame (Cipolla and Blake, 1992). Translation, along with divergence, curl, and deformation, will prove to be useful for describing facial expressions and are illustrated in Fig. 3. For example, when the motion of the eye regions is modeled as being affine, eye blinking can be detected as rapid deformation, divergence, and vertical translation in the eye region.

The affine model is not sufficient to capture the image motion of a human face when it occupies a significant portion of the field of view. A more appropriate model (which still provides a gross approximation to face motion) would assume that the face is a plane viewed under perspective projection. For small motions, the image motion of a rigid planar region of the scene can be described by the following eight-parameter model (Adiv, 1985; Waxman et al., 1987):

$$u(x, y) = a_0 + a_1 x + a_2 y + p_0 x^2 + p_1 xy, \qquad (6)$$

$$v(x, y) = a_3 + a_4 x + a_5 y + p_0 xy + p_1 y^2, \qquad (7)$$

where we have added two new terms p_0 and $p1$ to the affine model. These parameters roughly represent "yaw" and "pitch" deformations in the image plane respectively and are illustrated in Fig. 4. While we have experimented with more complex models of rigid face motion we have found that this planar assumption is both simple and expressive enough to robustly represent qualitative rigid facial motions in a variety of situations.

Non-rigid motions of facial features such as the eyebrows and mouth however are not well captured by the rigid affine or planar models. Deformable models such as snakes provide good tracking of these regions (Terzopoulos and Waters, 1993) but their distributed nature does not admit simple, intuitive, characterizations of the motions as we saw above but rather, necessitates additional analysis to extract meaningful

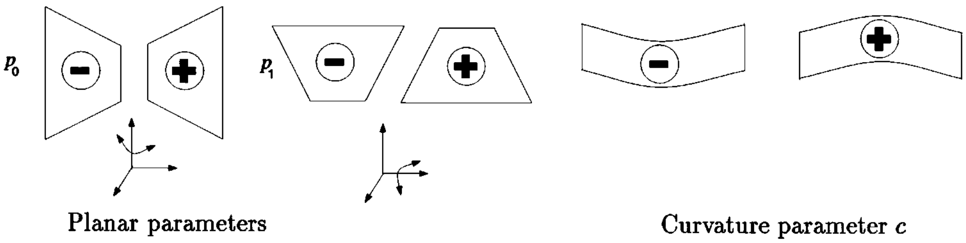


Figure 4. Additional parameters for planar motion and curvature.

descriptions of the motion. An alternative would be to use parameterized curve models for tracking (Blake and Isard, 1994). Deformable templates on the other hand (Yuille and Hallinan, 1992) encode information about shape but not motion. We wish to stay within the paradigm of using parametric models of image motion and so we augment the affine model to account for the primary form of curvature seen in mouths and eyebrows. We add a new parameter c to the affine model

$$u(x, y) = a_0 + a_1 x + a_2 y \tag{8}$$

$$v(x, y) = a_3 + a_4 x + a_5 y + c x^2 \tag{9}$$

where c encodes curvature and is illustrated in Fig. 4. As the experiments will demonstrate, this seven parameter model captures the essential image motion of the mouth and eyebrows.

Unfortunately this new curvature parameter is not invariant to head rotations. The curvature of the mouth and eyebrows should roughly be oriented with the principle axis of the face. To estimate the curvature with respect to the coordinate frame of the face we compute the orientation of the principle axis of the face and transform the images and features into the coordinate frame of the image plane. We then estimate the curvature and transform the features back into the coordinate frame of the face for the purpose of tracking.

3.2. Robust Regression

For convenience of notation we define

$$\mathbf{X}(\mathbf{x}) = \begin{bmatrix} 1 & x & y & 0 & 0 & 0 & x^2 & xy & 0 \\ 0 & 0 & 0 & 1 & x & y & xy & y^2 & x^2 \end{bmatrix} \tag{10}$$

$$\mathbf{A} = [a_0 \;\; a_1 \;\; a_2 \;\; a_3 \;\; a_4 \;\; a_5 \;\; 0 \;\; 0 \;\; 0]^T \tag{11}$$

$$\mathbf{P} = [a_0 \;\; a_1 \;\; a_2 \;\; a_3 \;\; a_4 \;\; a_5 \;\; p_0 \;\; p_1 \;\; 0]^T \tag{12}$$

$$\mathbf{C} = [a_0 \;\; a_1 \;\; a_2 \;\; a_3 \;\; a_4 \;\; a_5 \;\; 0 \;\; 0 \;\; c]^T \tag{13}$$

such that $\mathbf{u}(\mathbf{x}; \mathbf{A}) = \mathbf{X}(\mathbf{x})\mathbf{A}$, $\mathbf{u}(\mathbf{x}; \mathbf{P}) = \mathbf{X}(\mathbf{x})\mathbf{P}$, and $\mathbf{u}(\mathbf{x}; \mathbf{C}) = \mathbf{X}(\mathbf{x})\mathbf{C}$ represent, respectively, the affine, planar, and affine + curvature flow models described above.

Let f be the set of image points corresponding to the face region (excluding the non-rigid features), and P_f the planar motion parameters of these points. The brightness constancy assumption for the face states

$$I(\mathbf{x}, t) = I(\mathbf{x} - \mathbf{X}(\mathbf{x})\mathbf{P}_f, t + 1), \quad \forall \mathbf{x} \in f, \tag{14}$$

where I is the image brightness function and t represents time. Taking the Taylor series expansion of the right hand side, simplifying, and dropping terms above first order gives

$$\nabla I \cdot (\mathbf{X}(\mathbf{x})\mathbf{P}_f) + I_t = 0, \quad \forall \mathbf{x} \in f, \tag{15}$$

where $\nabla I = [I_x, I_y]$ and the subscripts indicate partial derivatives of image brightness with respect to the spatial dimensions and time.

To estimate the parameters $\mathbf{P}_f$ we minimize

$$\sum_{\mathbf{x} \in f} \rho(\nabla I \cdot (\mathbf{X}(\mathbf{x})\mathbf{P}_f) + I_t, \sigma), \tag{16}$$

for some error norm ρ where σ is a scale parameter. Since the face is neither a plane nor is it rigid it is important to take ρ to be a robust error norm which can cope with some percentage of gross errors or "outliers" (Hampel et al., 1986). For the experiments in this paper we take ρ to be

$$\rho(x, \sigma) = \frac{x^2}{\sigma^2 + x^2} \tag{17}$$

which is the robust error norm used by Geman-McClure (1987). As the magnitudes of residuals $\nabla I \cdot (\mathbf{X}(\mathbf{x})\mathbf{P}_f) + I_t$ grow beyond a point their influence on the solution begins to decrease and the value of $\rho(\cdot)$ approaches a constant. The value σ is a scale parameter that effects the point at which the influence of outliers begins to decrease.

Equation (16) is minimized using a simple gradient descent scheme with a continuation method that begins with a high value for σ and lowers it during the minimization (see (Black and Anandan, 1993, 1996) for details). The effect of this procedure is that initially no data are rejected as outliers then gradually the influence of outliers is reduced. To cope with large motions a coarse-to-fine strategy is used in which the motion is estimated at a coarse level then, at the next finer level, the image at time $t + 1$ is warped towards the image at time t using the current motion estimate. The motion parameters are refined at this level and the process continues until the finest level.

Once the face motion is estimated it is used to register the image at time $t + 1$ with the image at time t by warping the image at $t + 1$ back towards the image at t. Since the face is neither planar nor rigid this registration does not completely stabilize the two images. The residual motion is due either to the non-planar 3D shape of the head (its curvature and the nose for example) or the non-rigid motion of the facial features. We have observed that the planar model does a very good job of removing the rigid motion of the face and that the dominant residual motion is due to the motion of the facial features. The residual motion in the stabilized sequence is estimated using the appropriate motion model for that feature (i.e., affine or affine + curvature). Thus stablizing the face with respect to the planar approximation of its motion between two images allows the relative motions of the facial features to be estimated.

Note that the brightness constancy assumption used to estimate the image motion is often violated in practice due to changes in lighting, specular reflections, occlusion boundaries, etc. Robust regression has been shown to provide accurate motion estimates in a variety of situations in which the brightness constancy assumption in violated (Black and Anandan, 1996). When the brightness constancy assumption fails entirely or, when the image motion is not related to the true facial motion, a motion-based approach for facial expression recognition is likely to fail.

3.3. Tracking Facial Features

The estimated parametric motion of the face and facial features estimated between two frames is used to predict the location of the features in the next frame. The face and the eyes are simple quadrilaterals which are represented by the image locations of their four corners. Since a line on a plane remains a line under the planar motion $\mathbf{P}_f$ these regions remain quadrilaterals although the location of their four corners changes. We update the location $\mathbf{x}$ of each of the four corners of the face and eyes by applying the planar motion to get $\mathbf{X}(\mathbf{x})\mathbf{P}_f + \mathbf{x}$. Then the relative motion of the eyes locations is accounted for and the corners become $(\mathbf{X}(\mathbf{x})\mathbf{P})\mathbf{A}_{le} + \mathbf{x}$ and $(\mathbf{X}(\mathbf{x})\mathbf{P})\mathbf{A}_{re} + \mathbf{x}$ where le and re stand for the motions of the left and right eyes respectively. In updating the eye region we do not use the full affine model since when the eye blinks this would cause the tracked region to deform to the point where the eye region could no longer be tracked. Instead only the horizontal and vertical translation of the eye region is used to update its location relative to the face motion.

The curvature of the mouth and brows means that the simple updating of the corners is not sufficient for tracking. In our current implementation we use image masks to represent the regions of the image corresponding to the brows and the mouth. These masks are updated by warping them first by the planar face motion $\mathbf{P}_f$ and then by the motion of the individual features $\mathbf{C}_m$, $\mathbf{C}_{lb}$ and $\mathbf{C}_{rb}$ which correspond the mouth and the left and right brows respectively. This simple updating scheme works well in practice.

To reduce noise in the parameters we use a simple temporal filter. Let $\mathbf{P}_f^+$ be the filtered parameters of the face and $\mathbf{P}_f$ be the current estimate of the face parameters; then we update $\mathbf{P}_f^+$ as follows

$$\mathbf{P}_f^+ \leftarrow \frac{1}{2}(\mathbf{P}_f^+ + \mathbf{P}_f).$$

Exactly the same treatment is applied to the relative facial feature motions and these smoothed values are used for expression recognition. A more sophisticated Kalman filter could be used as in (Azarbayejani et al., 1993b) but this averaging scheme works well in our experiments and has the property of weighting current estimates more heavily than previous ones. This is an appropriate model for facial expressions which are typically of short duration. The one-sided nature of the filter causes a slight temporal shift in the parameters which has no significant effect on recognition.

4. Expression Recognition

The deformation and motion parameters described in the previous section can be used to derive mid- and

Table 1. The mid-level predicates derived from deformation and motion parameter estimates.

Parameter	Threshold	Derived predicates (mouth)
a_0	>0.25	Rightward
	<-0.25	Leftward
a_3	<-0.1	Upward
	>0.1	Downward
Div	>0.02	Expansion
	<-0.02	Contraction
Def	>0.005	Horizontal deformation
	<-0.005	Vertical deformation
Curl	>0.005	Clockwise rotation
	<-0.005	Counter clockwise rotation
c	<-0.0001	Curving upward (‘U’ like)
	>0.0001	Curving downward

Table 2. The mid-level predicates derived from deformation and motion parameter estimates as applied to head motion.

Parameter	Threshold	Derived predicates (head)
a_0	>0.5	Rightward
	<-0.5	Leftward
a_3	<-0.5	Upward
	>0.5	Downward
Div	>0.01	Expansion
	<-0.01	Contraction
Def	>0.01	Horizontal deformation
	<-0.01	Vertical deformation
Curl	>0.005	Clockwise rotation
	<-0.005	Counter clockwise rotation
p_0	<-0.00005	Rotate right about neck
	>0.00005	Rotate left about neck
p_1	<-0.00005	Rotate forward
	>0.00005	Rotate backward

high-level descriptions of facial actions; this section discusses these representations.

4.1. Mid-Level Representations

The parameters (such as translation and divergence) estimated for each feature are used to derive mid-level predicates that characterize the motion of the feature. The parameter values are first thresholded to filter out most of the small and noisy estimates. The mid-level representation describes the observed facial changes at each frame. Table 1 provides an example of the predicates for the ‘mouth.’ Similar tables were developed for the eyebrows and eyes using the same thresholds. These values are mainly dependent on the face size in the image (since it determines the image-motion measurement) and were set empirically from a few sequences.

The mid-level representation that describes the head motions is given in Table 2. The planar model of facial motion is primarily used to stabilize the head motion so that the relative motion of the features may be estimated. The motion of this plane also provides a qualitative description of the head motion. For example, we can qualitatively recover when the head is rotating or translating. To accurately recover the true 3D motion of the head would require a model more general than the planar assumption.

4.2. High-Level Representations

The high-level representation of facial actions (i.e., the facial expression recognition procedure) considers the temporal consistency of the mid-level predicates to minimize the effects of noise and inaccuracies in the motion and deformation models.

Following the temporal approach for recognition proposed in (Yacoob and Davis, 1994), we divide each facial expression into three temporal segments: the *beginning*, *apex* and *ending*. Figure 5 illustrates qualitatively the different aspects of detecting and segmenting a “smile.” In this figure the horizontal axis represents the time dimension (i.e., the image sequence), the axis perpendicular to the page represents each one of the parameters relevant to a ‘smile’ (i.e., a_3, *Div*, *Def*, and c) and the vertical axis represents the values of these parameters. This diagram is an abstraction to the progression of a smile, therefore the parameter values are not provided. Notice that Fig. 5 indicates that the change in parameter values might not occur at the same frames at either the beginning or ending of actions, but it is required that a significant overlap be detectable to label a set of frames with a “beginning of a smile” label, while the motions must terminate before a frame is labeled as an “apex” or an “ending”.

The detailed development of the smile model is as follows. The upward-outward motion of the mouth corners results in a negative curvature of the mouth (i.e., the curvature parameter c is negative). The horizontal and overall vertical stretching are manifested by positive divergence (Div) and deformation (Def). Finally, some overall upward translation is caused by the raising of the lower and upper lips due to the stretching of the mouth (a_3 is negative). Reversal of these

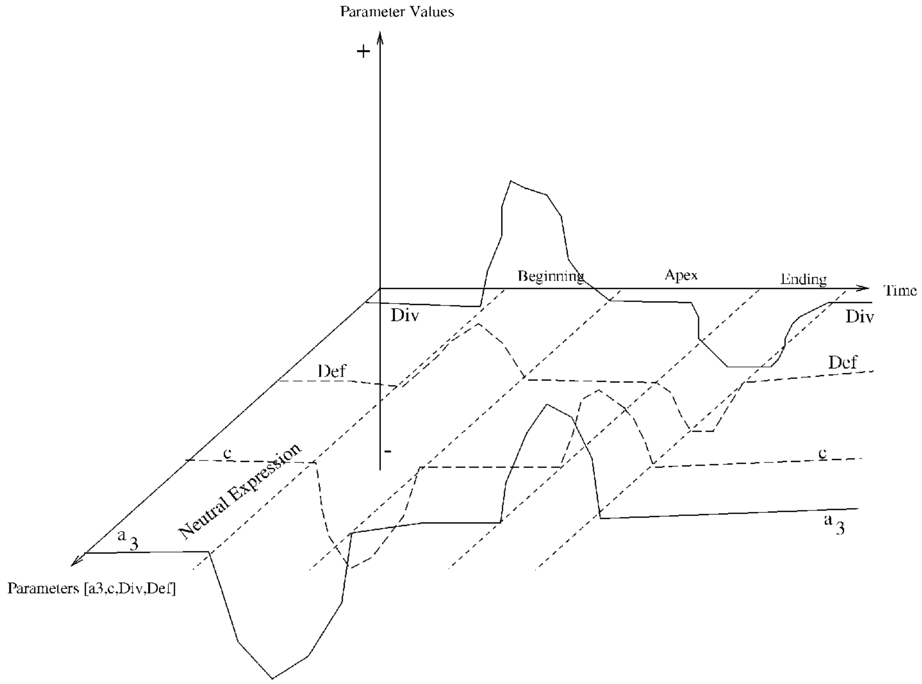


Figure 5. The temporal model of the “smile” expression.

motion parameters is observed during the ending of the expression.

The rules for detecting the beginning and ending of expressions are given in Table 3. These rules are applied to the predicates of the mid-level representation. Generally, a beginning/ending has to be detectable continuously over at least four consecutive frames for the action to be recognized. This temporal duration was determined empirically based on a video rate of 30 frames/second.

The high-level representation of head motion is currently limited to detecting backward and forward motions, right and left rotations around the neck and looming. These recognized motions are illustrated

Table 3. The rules for classifying facial expressions (B = beginning, E = ending).

Expr.	B/E	Satisfactory actions
Anger	B	Inward lowering of brows and mouth contraction
Anger	E	Outward raising of brows and mouth expansion
Disgust	B	Mouth horizontal expansion and lowering of brows
Disgust	E	Mouth contraction and raising of brows
Happiness	B	Upward curving of mouth and expansion or horizontal deformation
Happiness	E	Downward curving of mouth and contraction or horizontal deformation
Surprise	B	Raising brows and vertical expansion of mouth
Surprise	E	Lowering brows and vertical contraction of mouth
Sadness	B	Downward curving of mouth and upward-inward motion in inner parts of brows
Sadness	E	Upward curving of mouth and downward-outward motion in inner parts of brows
Fear	B	Expansion of mouth and raising-inwards inner parts of brows
Fear	E	Contraction of mouth and lowering inner parts of brows

in the experimental results found in the following sections along with the facial expressions that are detected.

4.3. Resolving Conflicts between Expressions

The system is designed to identify and recognize facial expressions from long video clips (i.e., clips including 3-6 expressions). We simplified the behavioral model of our subjects by asking them to display one emotion at a time, and include a short neutral state between expressions (we allowed, however, co-occurrences of ‘smile’ and ‘surprise’ since these were often displayed by our subjects). Since the six expression classifiers operate on the whole sequence independently, the system may create conflicting hypotheses.

Conflicts may arise when an ending of an expression is mistaken as the beginning of another expression. For example, the ‘anger’ recognition module may consider the lowering of the eyebrows during the ending of a ‘surprise’ expression as a beginning of an ‘anger’ expression. To resolve such conflicts, we employ a memory-based process that gives preference to the expression that started earlier.

5. Motion Estimation Results

The experiments in this section illustrate the rigid and non-rigid motion tracking while highlighting the information contained in the motion parameters that can be used for expression recognition. The first four experiments illustrate the relationship between the motion parameters and specific facial expressions (Happiness, Anger, Surprise, and Blinking) in the simple situation where the head is relatively stable. These experiments are followed by longer experiments on sequences of 100 or more images in which there is rigid head motion combined with the non-rigid motion of the facial expressions. The reader is referred to Figs. 3 and 4 in Section 3 to aid in understanding the parameters plotted in this section.

All parameters settings used in the motion estimation algorithm were exactly the same in all the experiments. In particular, for each pair of images, a Gaussian pyramid was constructed and 15 iterations of gradient descent were used at each level of the four levels in the coarse-to-fine strategy. The motion at one level is used to register the images before refining the estimate at the next finer level. The value of σ began at $15.0\sqrt{3}$, was lowered by a factor of 0.95 after each iteration, and was reset at each level in the pyramid. The initial regions for the head and the features were selected by hand and were automatically tracked thereafter. The initial segmentation need not be precise but the segmented regions should bound the features of interest.

5.1. Smile Experiment

The facial features tracked through the ‘smile’ image sequence are shown in Fig. 6. The figure shows the image regions corresponding to each of the features at a particular frame. As the ‘smile’ begins the curvature

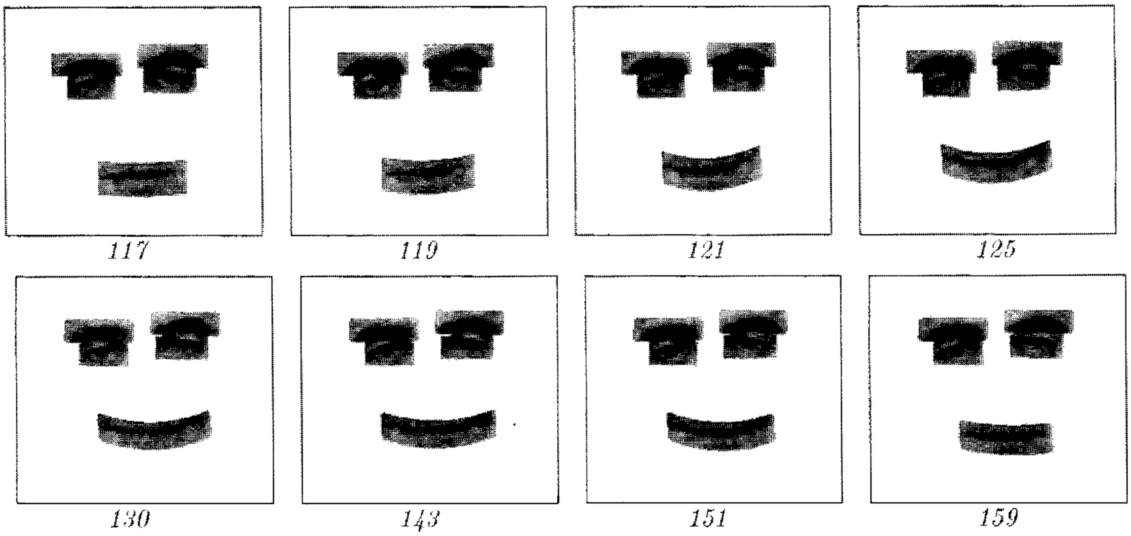


Figure 6. Smile experiment: facial expression tracking.

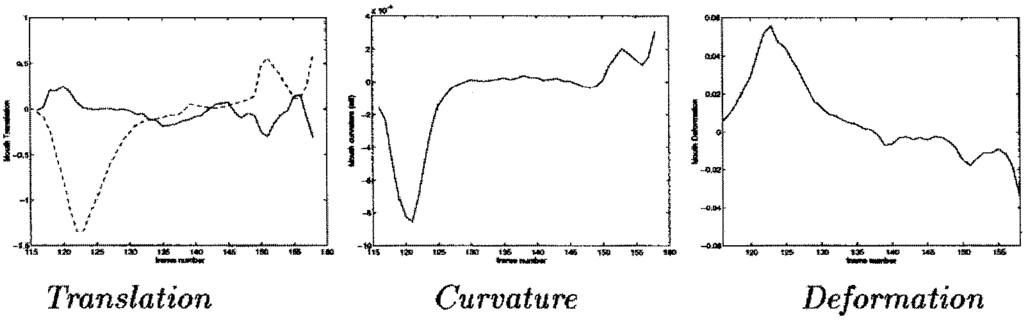


Figure 7. The smile mouth parameters. For translation, solid and dashed lines indicate horizontal and vertical motion respectively.

of the mouth is clearly visible; then this curvature is followed by an elongation of the mouth.

Figure 7 plots some of the parameters of the mouth motion. Note that the parameters represent image motion between frames not absolute motion. So, for example a plot of mouth "curvature" is a plot of the changing curvature between frames not the absolute "curvature" of the mouth. The left graph in Fig. 7 shows the horizontal (solid line) and vertical (dashed line) translation of the mouth. The negative vertical translation indicates that the mouth rises relative to the face during the initial phase of the 'smile' expression. The middle graph plots curvature of the mouth (parameter c) and clearly shows the negative curvature corresponding to an upwards bending motion at the initiation of the smile. The other significant cue to perceiving a 'smile' expression is a deformation of the mouth which resembles stretching in the horizontal direction; this is clearly visible in the plot of the mouth deformation $(a_1 - a_5)$ on the right in the figure.

5.2. Anger Experiment

An image sequence of an anger expression is shown in Fig. 8. The anger expression (Fig. 9) is characterized by an initial pursing (or flattening) of the lips then, in this case, a long slow downward curvature which ends abruptly around frame 150, after which the mouth curves and deforms back to the relaxed position. In addition to the mouth motion, the eyes and brows

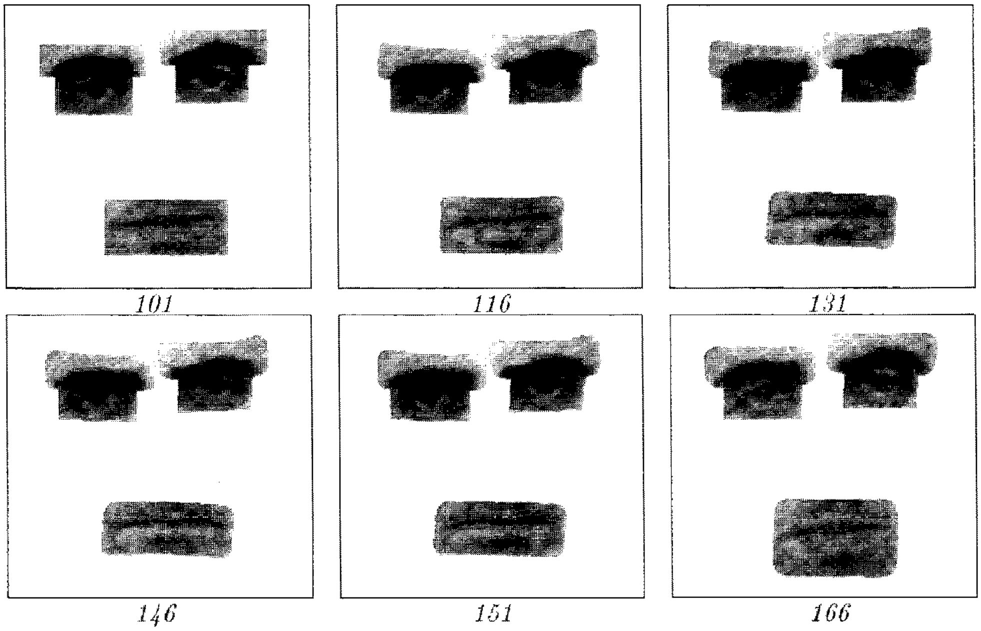


Figure 8. Anger Experiment: facial expression tracking. Features every 15 frames.

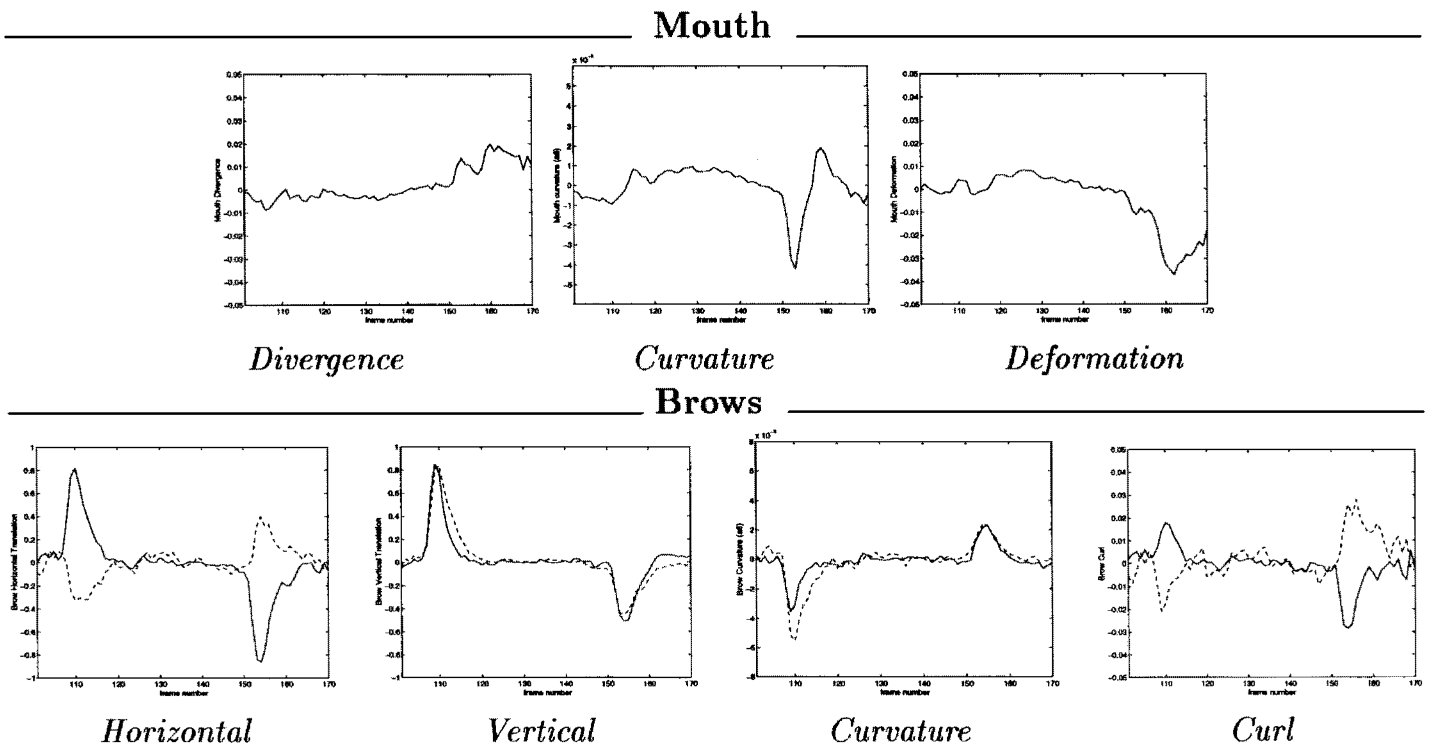


Figure 9. The Anger motion parameters; solid line indicates the right brow while the dashed line indicates the left brow.

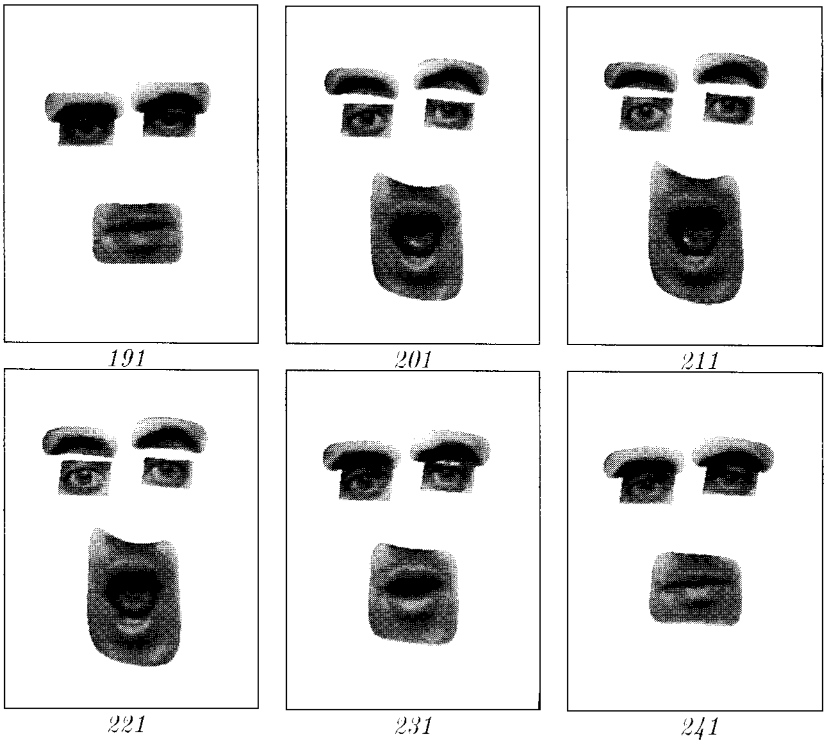


Figure 10. Surprise experiment: facial expression tracking. Features every 10 frames.

can play a significant role. Figure 9 shows how the brows move together and down while becoming flatter (negative curvature) during the initiation of the expression. The nasal edges of the brows also dip downwards causing opposite curl for the two brows. These motions are reversed on cessation of the expression.

5.3. *Surprise Experiment*

Features tracked during a "surprise" expression are shown in Fig. 10 and the significant parameters characterizing the expression are plotted in Fig. 11. During the initiation of the expression the mouth translates down, diverges, and deforms significantly. Simultaneously, the brows and eyes move upwards, the brows arch, and the eyes deform as they widen. The ending phase in this example shows a more gradual reversal of these parameters returning the face to the resting position.

5.4. *Blinking Experiment*

Figure 12 shows an image sequence in which the subject blinks twice. When the motion of the blinking eye is modeled as affine between two frames, the blinking is readily apparent in the plots of vertical translation, divergence, and deformation. Figure 13 shows the rapid onset of a blink around frame 245. The blink subsides more gradually than it began and then another blink begins around frame 260 and likewise reverses more gradually. Notice that the tracked eye regions are not affected by the blinking action.

5.5. *Looming Experiment*

The image sequence in Fig. 14 illustrates facial expressions (smiling and surprise) in conjunction with rigid head motion (in this case looming). The figure plots the regions corresponding to the face and the facial features tracked across the image sequence. The parameters describing the planar motion of the face are plotted in Fig. 15 where the divergence due to the looming motion of the head is clearly visible in the plot of divergence. Analyzing the plots of the facial features in Fig. 16 reveals that a smile expression begins around frame 125 with an increase in mouth curvature followed by a deformation of the mouth. The curvature decreases between frames 175 and 185 and

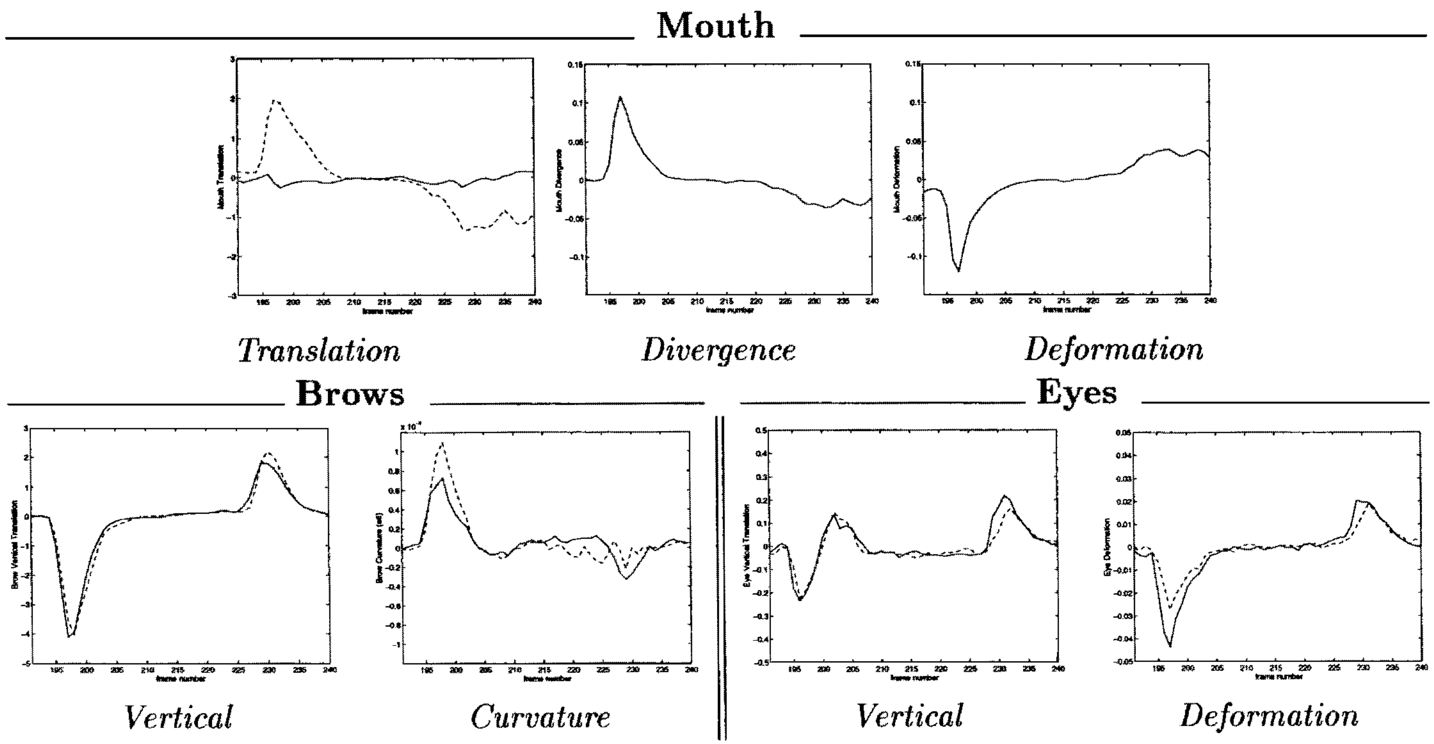


Figure 11. The surprise parameters. For the mouth translation, the solid line indicates horizontal motion while the dashed line indicates vertical motion. For the eye and brows, the solid and dashed lines indicate left and right respectively.

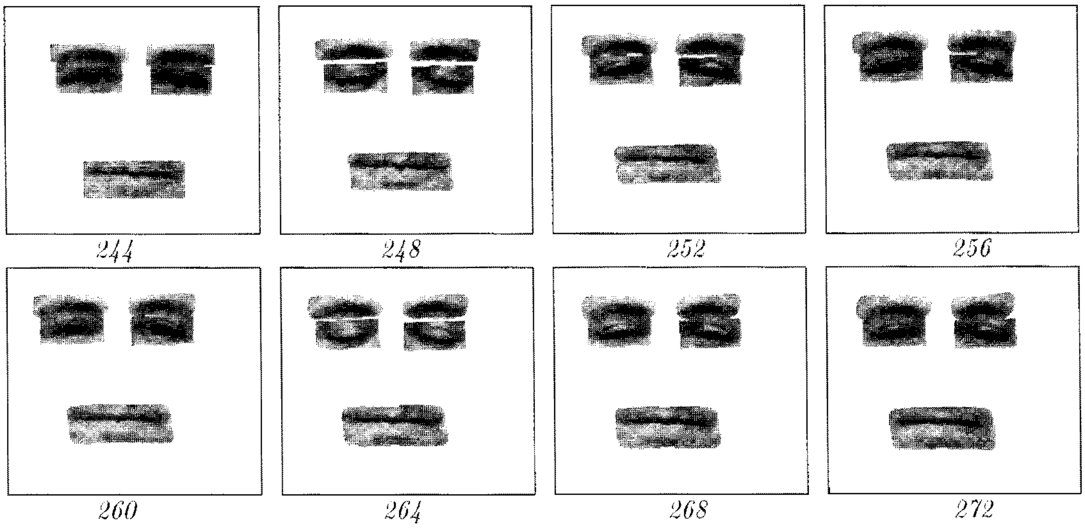


Figure 12. Blinking experiment: facial feature tracking. Features every 4 frames.

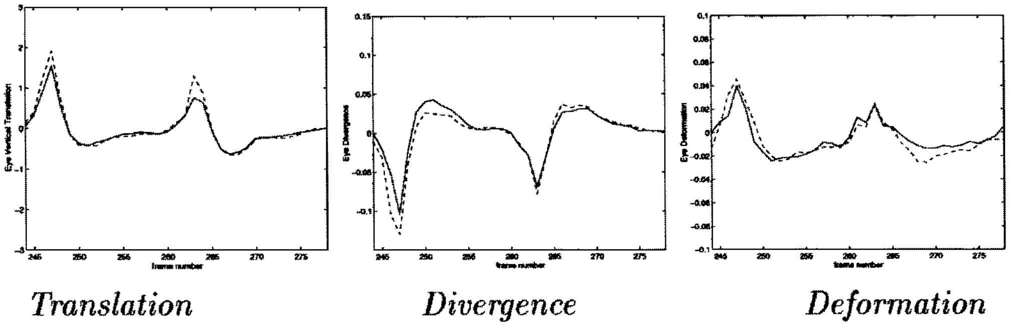


Figure 13. The blinking experiment; motion parameters for the eyes. Left eye = solid line. Right eye = dashed line.

then a surprise expression begins around frame 220 with vertical eyebrow motion, brow arching, and mouth deformation.

5.6. Rotation Experiment

Figure 17 illustrates a sequence with more complex rigid face motion due to rapid head rotations. In addition to the rapid motion, the sequence contains sections of motion blur, loss of focus, and saturation. Despite the low quality of the sequence the tracking of the head and features is robust.

In Fig. 18 the plot of curl shows that the face rotates clockwise in the image plane then rotates counterclockwise, pauses briefly, and continues the counterclockwise motion. The plot of p_0 (solid line) indicates that the head is rotating about the neck axis roughly in conjunction with the curl. To a lesser degree the head pitches fore and aft (p_1—dashed line) and looms forwards and back.

The facial features are consistently tracked despite the large image motions and the recovered parameters in Fig. 19 indicate that two surprise expressions take place during the sequence as characterized by the eyebrow motion and mouth deformation (see the surprise example with a static head for comparison.)

6. Recognition Results

We carried out a large set of experiments to verify and evaluate the performance of the recognition procedure proposed in this paper. The first set of experiments focuses on the expressions of forty subjects who were asked to perform expressions in front of a video-camera. The second set of experiments involved

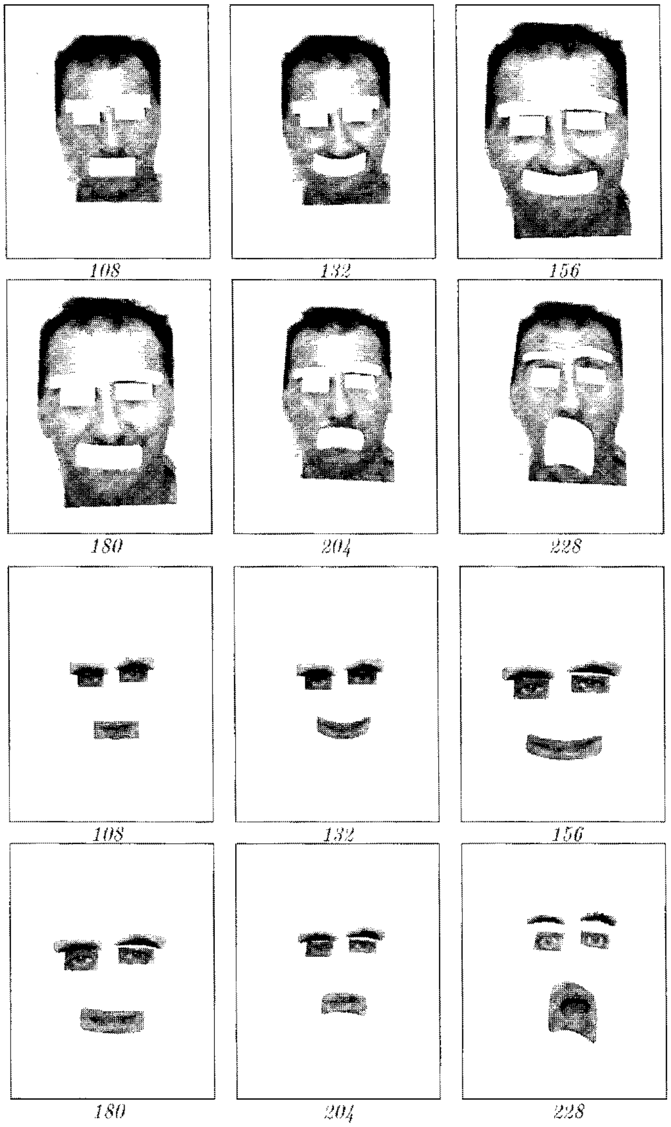


Figure 14. Looming experiment. Facial expression tracking with rigid head motion (every 24 frames).

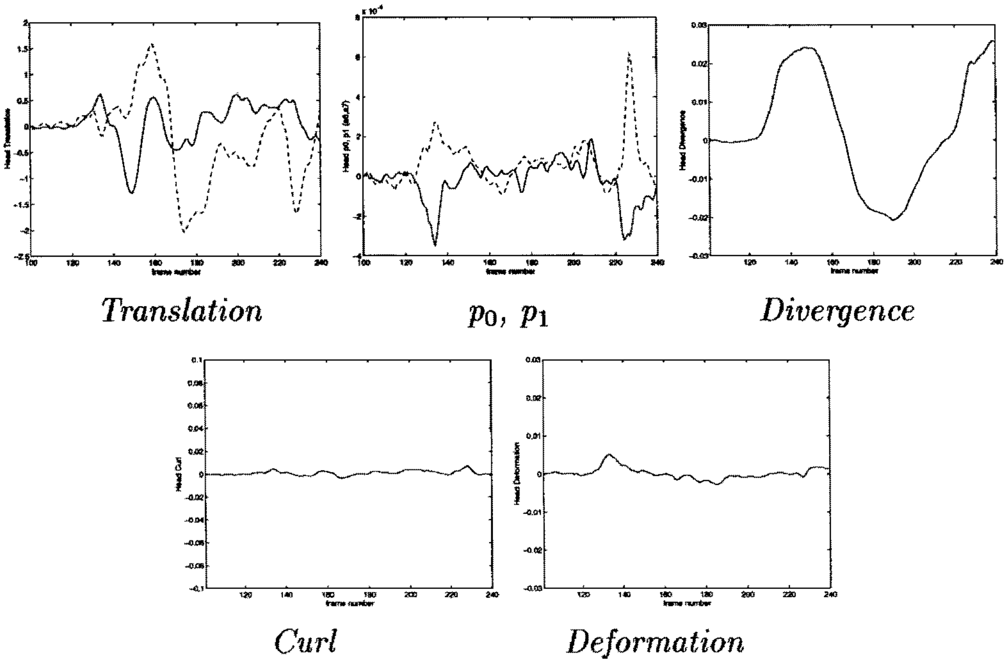
digitizing video-clips from television and movies. In the rest of this section we discuss the methodology used in acquiring the data and provide detailed statistics and analysis of the results.

6.1. *Methodology*

There are both technical difficulties in collecting data sets of sufficient spatial and temporal resolution and challenges in designing and interpreting experiments that evaluate the ability of a system to recognize expressions. Technically,

- image sequences should be sampled at 30 Hz (or more) to minimize the magnitudes of rigid and non-rigid motions between consecutive images,
- image resolution should allow the facial features to be of sufficient size to facilitate tracking and motion estimation, and
- the amount of data that needs to be captured and processed is large (on the order of 10 MBytes per second) so that only expressions of short duration can be captured.

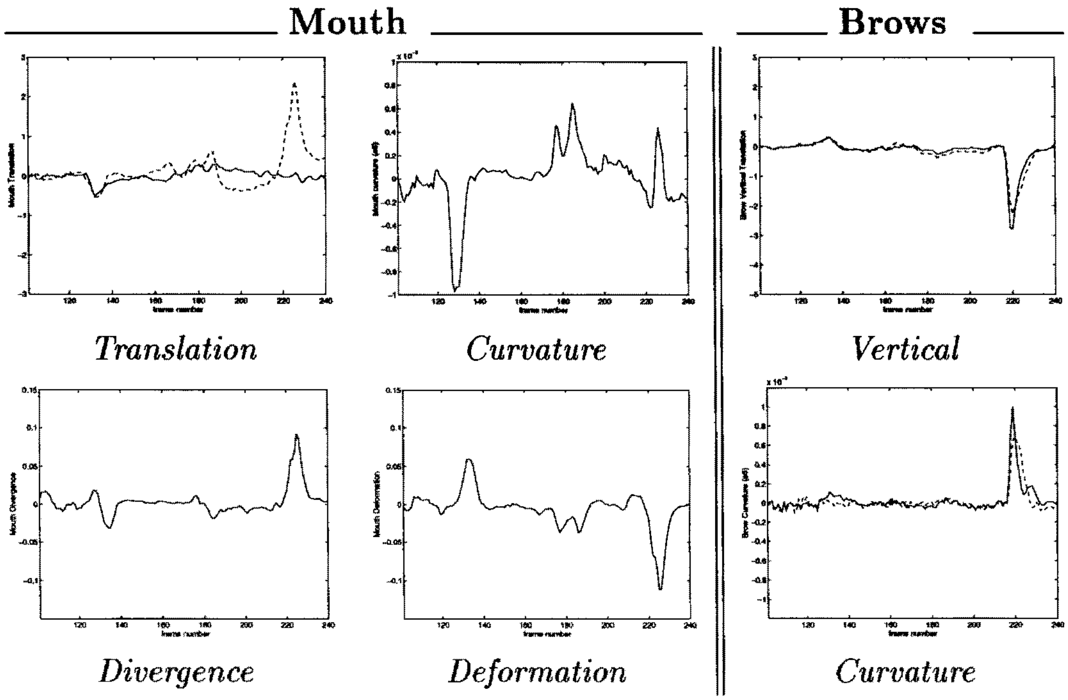
Challenges associated with experimental designs include



Translation *p_0, p_1* *Divergence*

Curl *Deformation*

Figure 15. The looming face motion parameters. Translation: solid = horizontal, dashed = vertical. Quadratic terms: solid = p_0, dashed = p_1.



Mouth **Brows**

Translation *Curvature* *Vertical*

Divergence *Deformation* *Curvature*

Figure 16. The looming sequence. Mouth translation: solid and dashed lines indicate horizontal and vertical motion respectively. For the brows, the solid and dashed lines indicate left and right brows respectively.

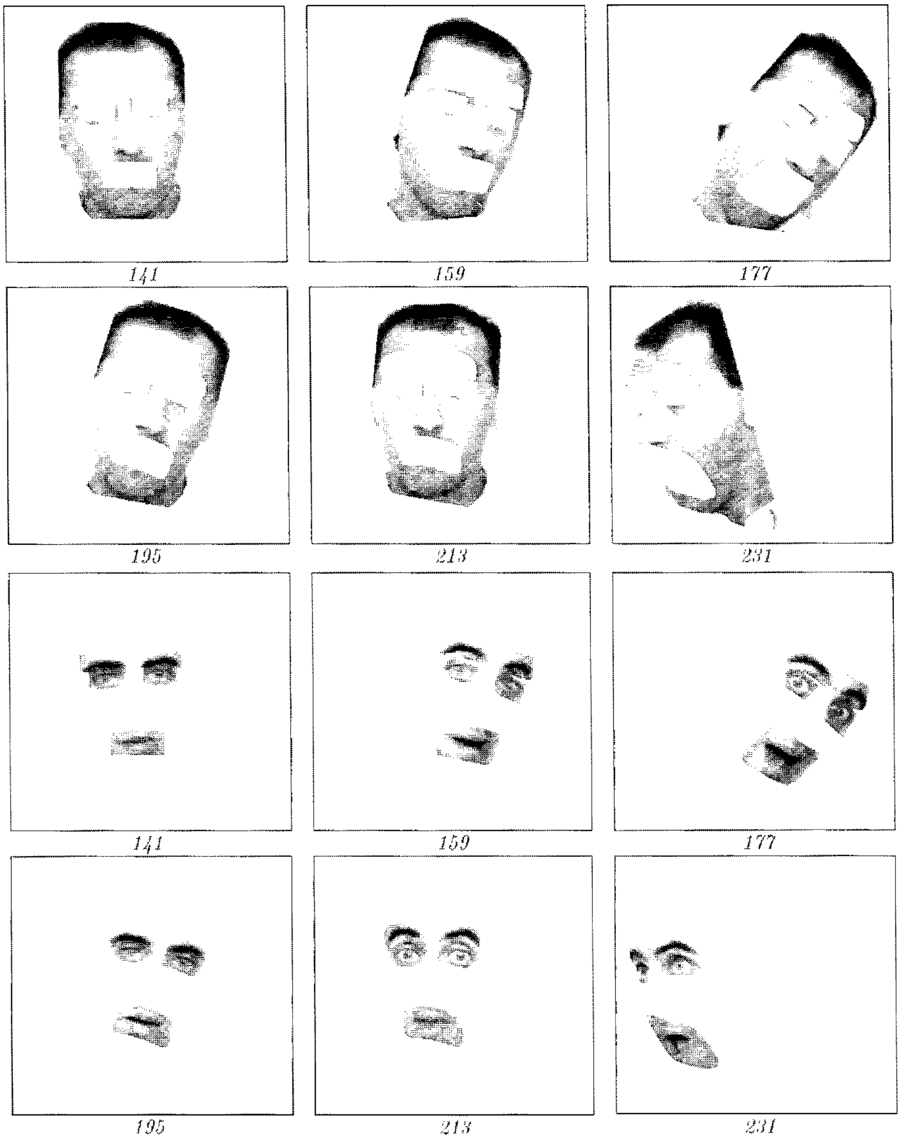


Figure 17. Rotation experiment. Rigid head tracking, every 18th frame.

- expressions of emotions are hard to elicit in an artificial environment where people are not experiencing the typical associated stimuli (see (Ekman, 1982) for a discussion about how such experiments are conducted in psychology),
- determining what expression is "actually" being displayed is difficult because "different" expressions may appear quite similar leading to variation in human recognition of expressions.

In an attempt to maintain some consistency in deciding which expression is actually being displayed, we used the the cues identified by psychological studies (Bassili, 1979; Ekman and Friesen, 1975) to determine the "ground truth" expressions.

We provide two sets of experiments. In the first, we recorded tens of expressions of forty subjects (having varied race, culture, and appearance) displaying their own choice of expressions. Our experimental subjects were asked to display emotional expressions without additional directions (in fact, we asked the subjects to choose any subset of expressions and display the expressions in any order and as naturally as they possibly could). The expressions 'fear' and 'sadness' were hard

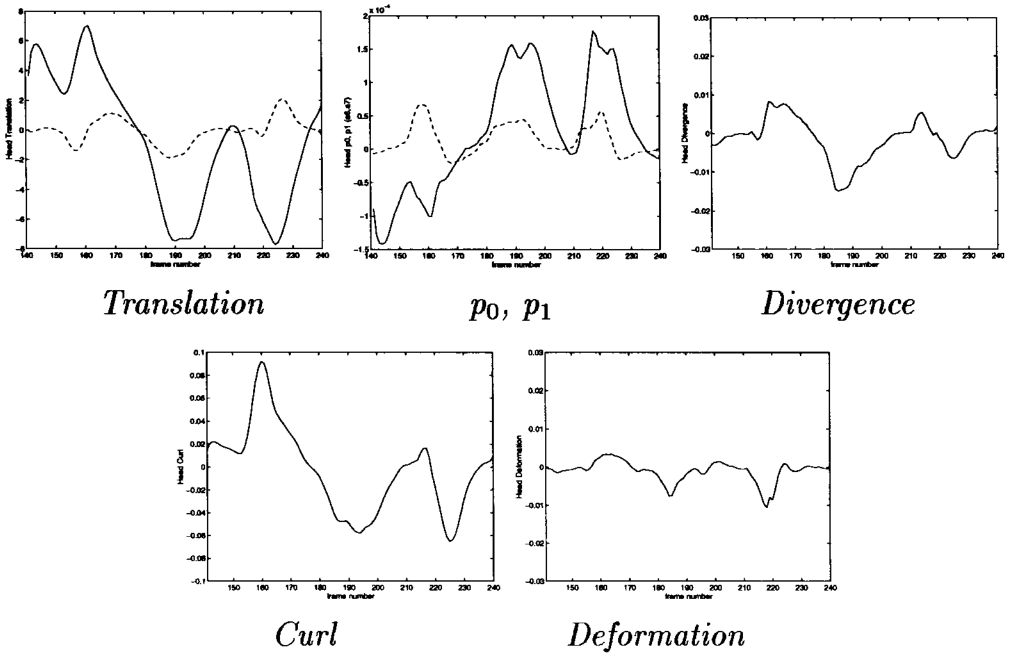


Figure 18. The rotate sequence face motion parameters. Translation: solid = horizontal, dashed = vertical. Quadratic terms: solid = p_0, dashed = p_1.

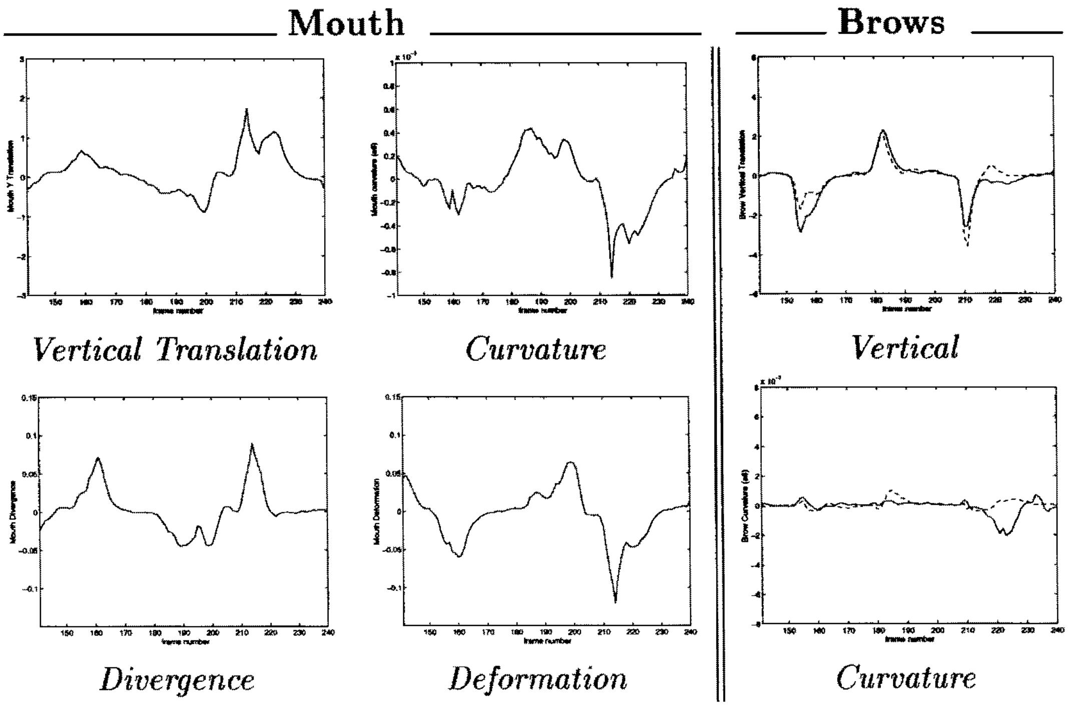


Figure 19. The rotate sequence. For the brows, the solid and dashed lines indicate left and right brows respectively.

to elicit compared to 'happiness,' 'surprise,' 'disgust' and 'anger.' The subjects were also asked to move their heads; but avoid a profile view. In the second set of experiments, we digitized tens of video-clips recorded from talk shows, news, and movies. TV broadcasting, reception, video-recording and digitization make the data quite noisy.

While we asked our laboratory subjects not to speak while conveying expressions, the video-recordings only occasionally include speechless-expressions. The occurrence of speech affects the interpretations of the mouth motions since some motions due to speech may appear to be due to expressions. Therefore, we chose clips that are closely associated with emotional expressions (emotional behavior on talk-shows, 'smiles' as a substitute for 'happiness,' etc.). The set of expressions recorded is dominated by 'smiles.'

6.2. Results of Laboratory Set-Up

Figure 20 shows some of the forty subjects who participated in our study; from these subjects we collected a database of 70 image sequences containing a total of 145 expressions. Each sequence is about 9 seconds long and contains 1–3 expressions. Images are 560×420 pixels (taken at 30 Hz).

The results of the experiments are summarized in Table 4. The first column breaks the expressions in the sequence into the six basic categories according to the "ground truth." The second column in the table notes the number of occurrences of each of the expression types. "Correct" indicates that the expression was of type x and was judged to be of type x. "Insertions," or false positives, occurred when the expression was judged to be of type x but the ground truth indicated that the expression was neutral. Insertions could also occur when the system answered that the expression was of type x in the midst of another expression. Insertions were primarily due to motion and tracking inaccuracies. For example, rigid head motion might be mistaken for a smile or a smile might be recognized in the middle of a surprise expression. "Deletions" indicate that the expression was of type x but that it was not noticed by the system. "Substitutions" occurred when the expression was of type x but was judged to be of some other type by the system; that is, a confusion of one expression with another. The sum of the deletions and substitutions gives the total number of misclassified expressions.

Figure 20. Twenty of the forty participants in the experiments.

Table 4. Facial expression recognition results on forty subjects.

Expression	Occurrences	Correct	Insertions	Deletions	Substitutions	Success rate	Accuracy
Happiness	61	58	7	1	2	95%	84%
Surprise	35	32	—	3	—	91%	91%
Anger	20	18	—	2	—	90%	90%
Disgust	15	14	1	1	—	93%	87%
Fear	6	5	—	—	1	83%	83%
Sadness	8	8	—	—	—	100%	100%
Total	145	135	8	7	3	93%	88%

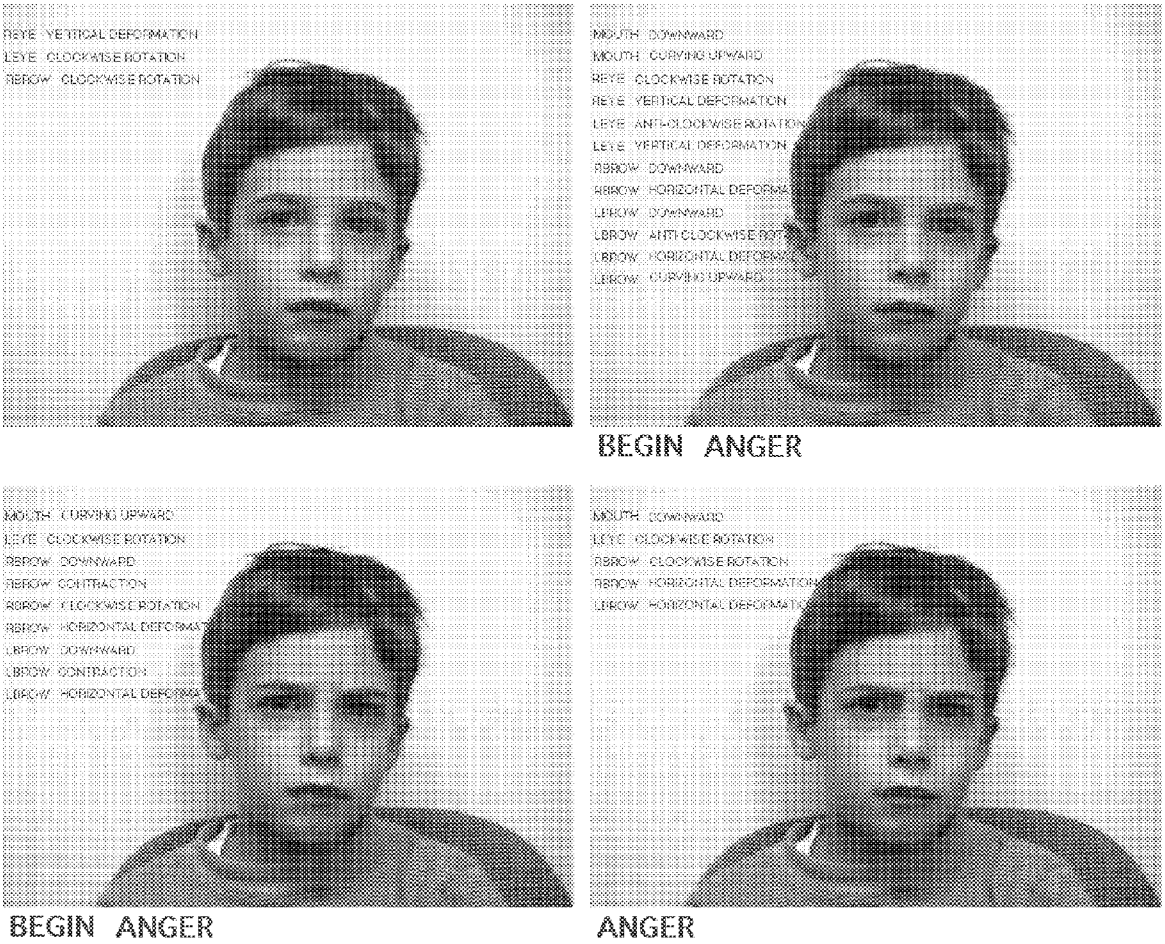


Figure 21. Four frames (four frames apart) of the beginning of an 'anger' expression displayed by a six year old boy.

The performance of the approach is judged in two ways. The "Success Rate" and "Accuracy" are defined as:

$$\text{Success: } \frac{\text{Correct}}{\text{Occurrences}},$$

$$\text{Accuracy: } \frac{\text{Correct} - \text{Insertions}}{\text{Occurrences}}.$$

The overall success rate for the system was 93% while the accuracy was 89%.

The dynamic nature of facial expressions makes it difficult to demonstrate the experiments in print. Therefore, we provide selected images that will, hopefully, convey our results. Figure 21 shows four frames (taken as every fourth frame from the sequence) of the

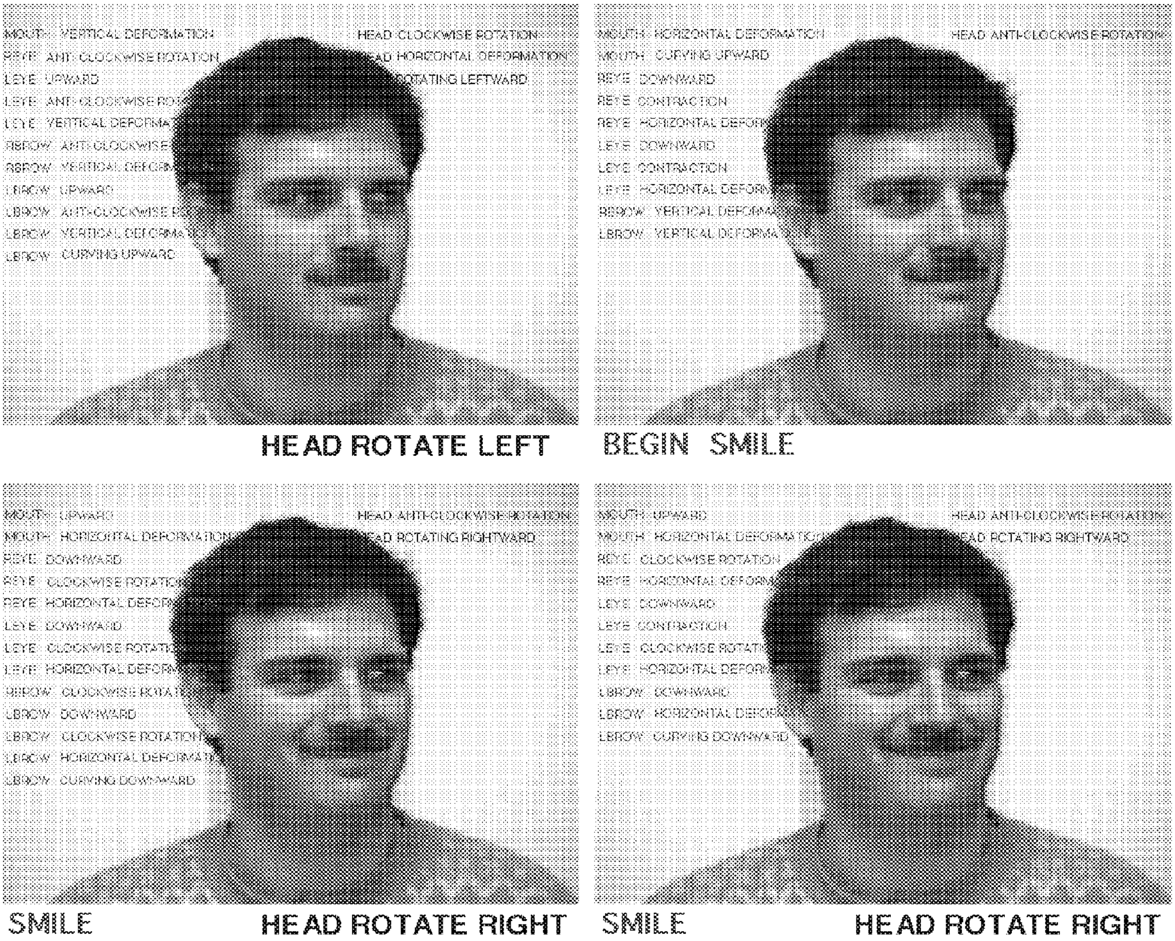


Figure 22. Four frames (four frames apart) of the beginning of a 'smile' expression.

beginning of an 'anger' expression of a six year old boy. The text that appears on the left side of each image represents the mid-level predicates of the facial deformations, and the text that appears on the right side represents the mid-level predicates of the head motion. The text below each image displays the high-level description of the facial deformations and the head motions. Figure 22 shows the beginning of a 'smile' expression while the head is rotating initially leftward and then rightward.

6.3. *Results of Video-Clips*

Figure 23 shows a representative sample of the 36 video-clips that were collected from TV talk shows, news, movies, and commercials. Table 5 shows the details of our results on these 36 video-clips.

Figure 24 shows four frames (taken as every fourth frame from the sequence) of the beginning of a "surprise" expression of a TV-show host. Note that the expression is not a classical "surprise" expression; the eyebrows and eyes deform in the appropriate way but the mouth does not. The system classifies every expression as one of the known expressions or no expression. Since some expression is occurring the system chooses the most likely one which, in this case, is "surprise".

The analysis of another example clip is provided in Fig. 25. This figure shows four frames taken from the movie "Amadeus" in which Mozart displays 'fear.'

6.4. *Computational Cost*

With the exception of the rigid and non-rigid motion estimation computations, the algorithms described

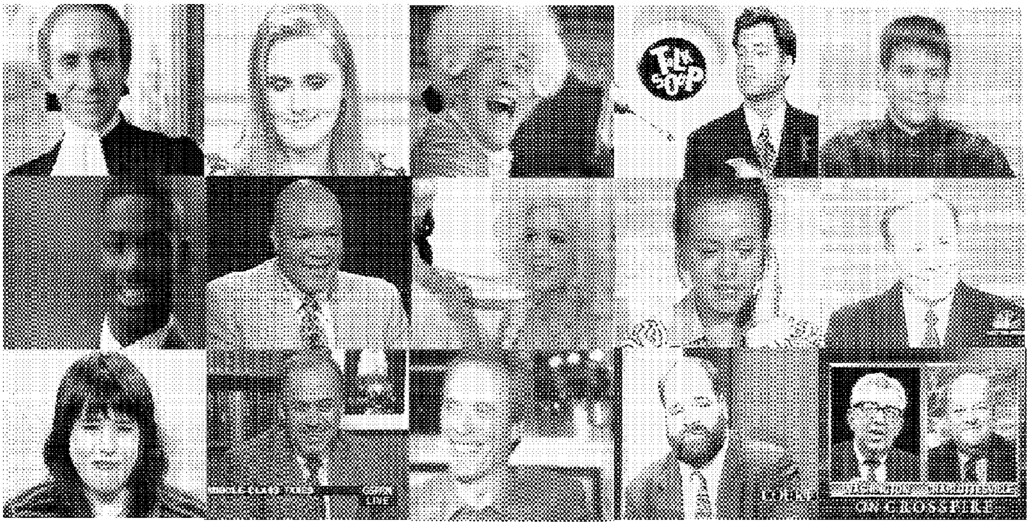


Figure 23. Representative sample of still frames from the 36 video-clips.

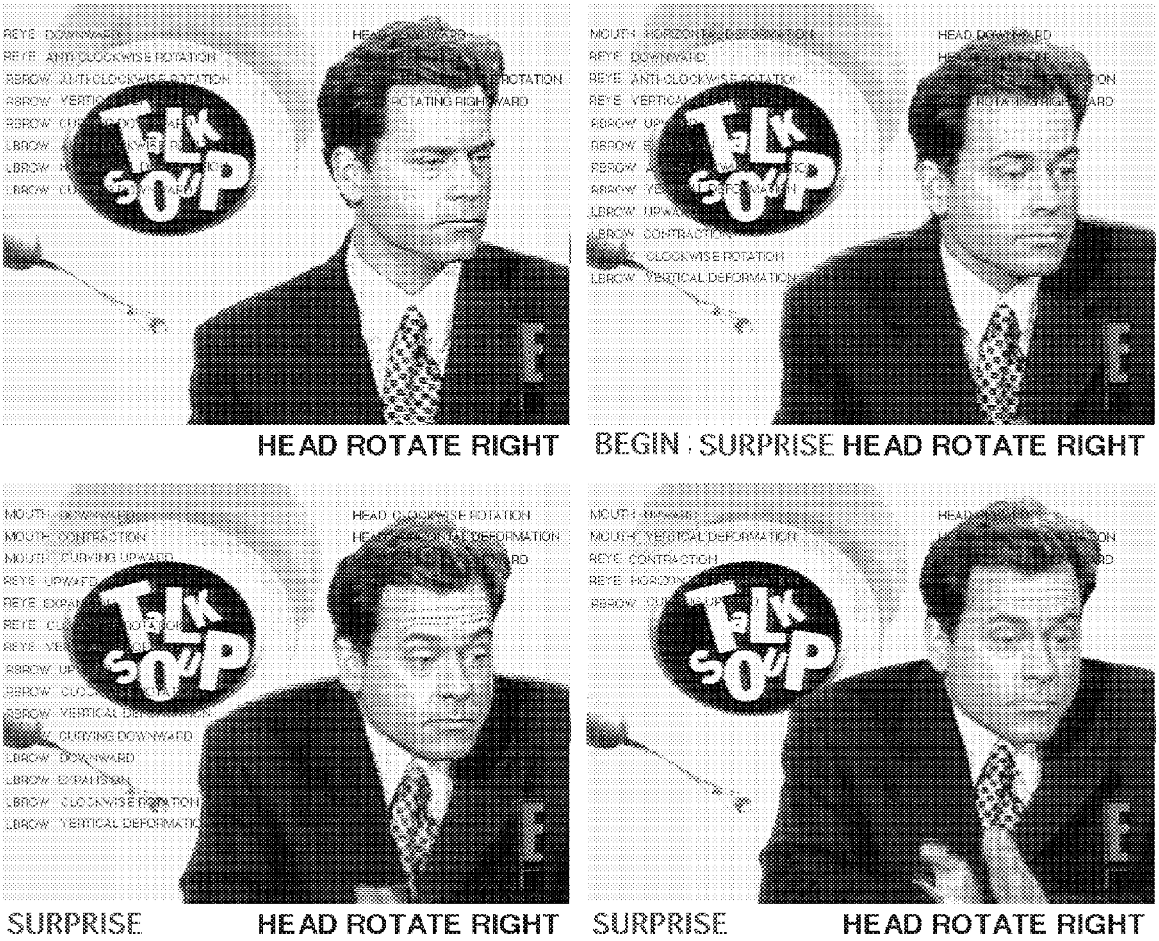


Figure 24. Four frames (four frames apart) of the beginning of a 'surprise' expression.

Table 5. Facial expression recognition results on 36 video-clips.

Expression	Occurrences	Correct	Insertions	Deletions	Substitutions	Success rate	Accuracy
Happiness	37	35	4	2	—	95%	84%
Surprise	7	6	1	1	—	86%	71%
Anger	5	4	—	—	1	80%	80%
Disgust	4	2	2	2	—	50%	0%
Fear	1	1	—	—	—	100%	100%
Sadness	5	3	1	2	—	60%	40%
Total	59	51	8	7	1	86%	73%

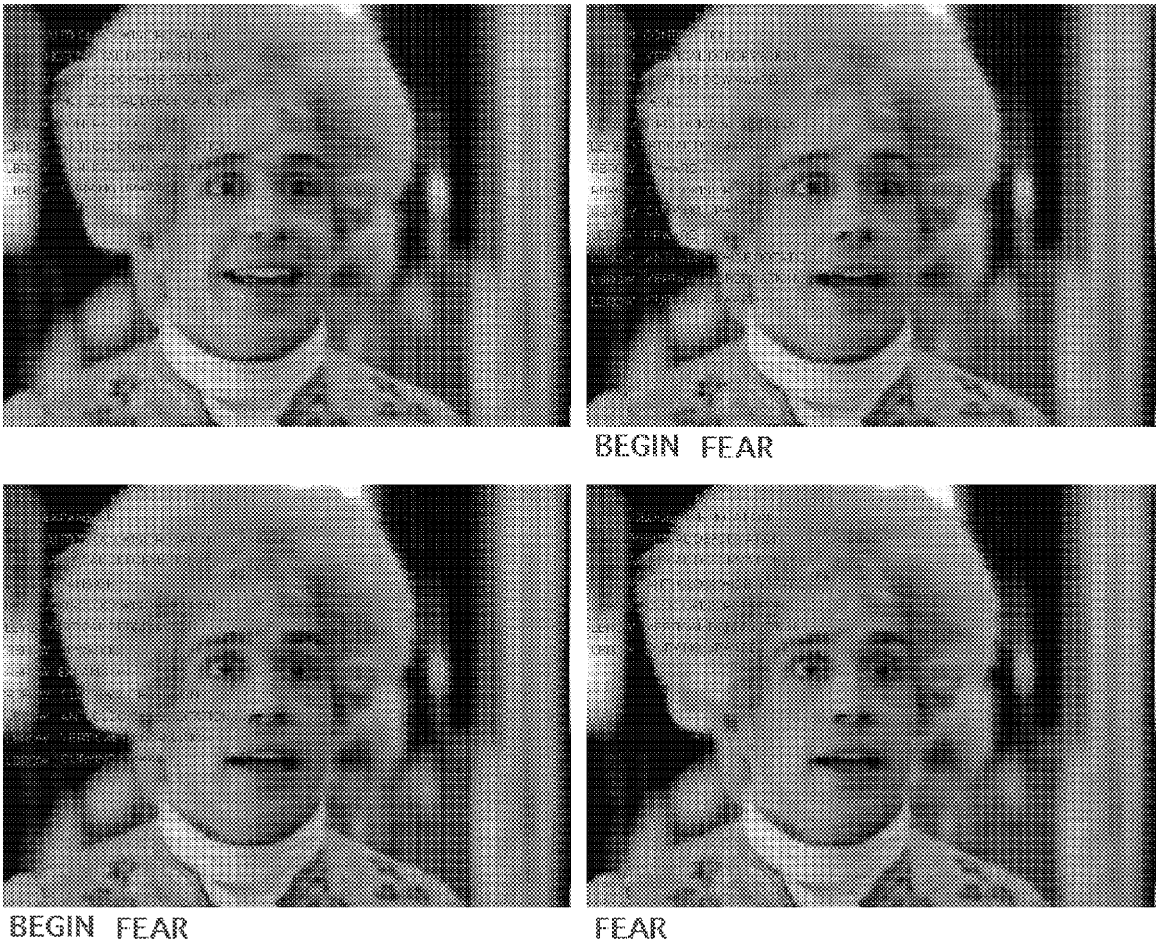


Figure 25. Four frames (four frames apart) of the beginning of a 'fear' expression.

operate at near frame-rates. The recovery of deformation and motion parameters requires about 2 minutes per frame on a Digital *Alpha* 3000 at full image resolution. We are currently studying possible speedups for the computations using an incremental estimation approach.

6.5. *Observations from Experimental Work*

Our experiments illustrate the wide variety of situations for which our approach works and the robustness of the tracking and recognition procedures. Below we explore some of the limitations of the current work.

- The planar assumption used to model the face can lead to quantitative tracking errors for subjects with very spherical faces. Additionally, we want to analyze profile views of the head, but cannot currently track the motion of the head between profile to frontal views.
- Various natural mouth motions (lip biting, and even simple opening) are sometimes mistakingly identified as 'smiles'. Also, speech remains a considerable source of false alarms (usually, openings are confused with 'smiles').
- Shape information is not directly integrated into our analysis, yet the use of shape can considerably improve the accuracy of the system at all levels. At the low level, tracking and feature deformation can benefit from feedback from a shape recognition module. At the high level the shape created by contours around the features can disambiguate expressions.

In addition, some technical observations:

- Accuracy of the motion estimation decreases whenever faces occupy too small a portion of the image (about 1/8 of the NTSC standard). This limits somewhat the cases where the facial expression analysis is applicable.
- Interlacing in broadcasting and very rapid body motions produce images that do not satisfy the basic assumption of brightness conservation inherent in the motion estimation. The standard approach of processing only one of the interlaced fields at a time results in reduced image sizes and hence smaller features whose motion cannot be estimated as reliably.

7. Conclusion and Future Research

In this paper we proposed local parameterized models of image motion that can recover the rigid and non-rigid facial motions that are an integral part of human behavior. The motions of facial features are modeled locally to allow for accurate recovery of their deformations. A robust optical flow algorithm is developed to the recover the motion of a person's head and the relative motions of their facial features. In a series of experiments we have demonstrated how these parameterized models of optical flow provide a concise description of the human motion in terms of a few parameters and how these parameters can be used to recognize human expressions. The paper has presented a facial-expression recognition strategy based on these motion models and we have illustrated the effectiveness of the approach for recognizing facial expressions. Extensive experimentation with many subjects in natural situations, including television clips, indicates that expression recognition from motion can be achieved accurately even in the presence of significant head motion.

Our ongoing work is focused on the modeling of coincident speech and facial expressions, more accurate recovery of 3D head motion, and dealing with a richer set of facial expressions. In particular we are exploring issues related to the integration of our motion-based approach with shape-based approaches. We are also pursuing the application of facial expression tracking and recognition to user interfaces.

Acknowledgments

We are thankful to Prof. Larry Davis for his valuable comments and suggestions. Without the wonderful cooperation of our forty volunteers in subjecting themselves to the intrusive experiments on their facial expressions this work would not have been possible. To them we are grateful.

References

Adiv, G. 1985. Determining three-dimensional motion and structure from optical flow generated by several moving objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-7(4):384–401.

Azarbayejani, A., Horowitz, B., and Pentland, A. 1993a. Recursive estimation of structure and motion using relative orientation constraints. In *Proc. Computer Vision and Pattern Recognition, CVPR-93*, New York, pp. 294–299.

Azarbayejani, A., Starner, T., Horowitz, B., and Pentland, A. 1993b. Visually controled graphics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(6):602–604.

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

Bergen, J.R., Anandan, P., Hanna, K.J., and Hingorani, R. 1992. Hierarchical model-based motion estimation. In *Proc. of Second European Conference on Computer Vision, ECCV-92*, G. Sandini (Ed.), Springer-Verlag, volume 588 of LNCS-Series, pp. 237–252.

Beymer, D., Shashua, A., and Poggio, T. 1993. Example based image analysis and synthesis. Technical Report A.I. Memo No. 1431, MIT.

Black, M.J. and Anandan, P. 1993. A framework for the robust estimation of optical flow. In *Proc. Int. Conf. on Computer Vision, ICCV-93*, Berlin, Germany, pp. 231–236.

Black, M.J. and Jepson, A. 1994. Estimating multiple independent motions in segmented images using parametric models with local deformations. In *Proceedings of the Workshop on Motion of Non-rigid and Articulated Objects*, Austin, Texas, pp. 220–227.

Black, M.J. and Anandan, P. 1996. The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. *Computer Vision and Image Understanding*, 63(1):75–104.

Blake, A. and Isard, M. 1994. 3D position, attitude and shape input using video tracking of hands and lips. In *Proceedings of SIGGRAPH 94*, pp. 185–192.

Chow, G. and Li, X. 1993. Towards a system for automatic facial feature detection. *Pattern Recognition*, 26(12):1739–1755.

Cipolla, R. and Blake, A. 1992. Surface orientation and time to contact from image divergence and deformation. In *Proc. of Second European Conference on Computer Vision, ECCV-92*, G. Sandini (Ed.), Springer-Verlag, volume 588 of LNCS-Series, pp. 187–202.

Ekman, P. 1992. Facial expressions of emotion: An old controversy and new findings. *Philosophical Transactions of the Royal Society of London*, B(335):63–69.

Ekman, P. and Friesen, W. 1975. *Unmasking the Face*. Prentice Hall.

Ekman, P. (Ed.) 1982. *Emotion in the Human Face*. Cambridge University Press.

Essa, I.A. and Pentland, A. 1994. A vision system for observing and extracting facial action parameters. In *Proc. Computer Vision and Pattern Recognition, CVPR-94*, Seattle, WA, pp. 76–83.

Essa, I., Darrell, T., and Pentland, A. 1994. Tracking facial motion. In *Proceedings of the Workshop on Motion of Non-rigid and Articulated Objects*, Austin, Texas, pp. 36–42.

Geman, S. and McClure, D.E. 1987. Statistical methods for tomographic image reconstruction. *Bulletin of the International Statistical Institute*, LII-4:5–21.

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., and Stahel, W.A. 1986. *Robust Statistics: The Approach Based on Influence Functions*. John Wiley and Sons: New York, NY.

Kass, M., Witkin, A., and Terzopoulos, D. 1987. Snakes: Active contour models. In *Proc. First International Conference on Computer Vision*, pp. 259–268

Koenderink, J.J. and van Doorn, A.J. 1975. Invariant properties of the motion parallax field due to the movement of rigid bodies relative to an observer. *Optica Acta*, 22(9):773–791.

Li, H., Roivainen, P., and Forcheimer, R. 1993. 3-D motion estimation in model-based facial image coding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(6):545–555.

Mase, K. 1991. Recognition of facial expression from optical flow. *IEICE Transactions*, E 74:3474–3483.

Rosenblum, M., Yacoob, Y., and Davis, L.S. 1994. Human emotion recognition from motion using a radial basis function network architecture. In *Proceedings of the Workshop on Motion of Non-rigid and Articulated Objects*, Austin, Texas.

Terzopoulos, D. and Waters, K. 1993. Analysis and synthesis of facial image sequences using physical and anatomical models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(6):569–579.

Toelg, S. and Poggio, T. 1994. Towards an example-based image compression architecture for video-conferencing. Technical Report CAR-TR-723, Center for Automation Research, U. of Maryland.

Waxman, A.M., Kamgar-Parsi, B., and Subbarao, M. 1987. Closeform solutions to image flow equations. In *Proc. Int. Conf. on Computer Vision, ICCV-87*, London, England, pp. 12–24.

Yacoob, Y. and Davis, L.S. 1993. Labeling of human face components from range data. In *Proc. Computer Vision and Pattern Recognition, CVPR-94*, New York, NY, pp. 592–593.

Yacoob, Y. and Davis, L.S. 1994. Computing spatio-temporal representations of human faces. In *Proc. Computer Vision and Pattern Recognition, CVPR-94*, Seattle, WA, pp. 70–75.

Young, A.W. and Ellis, H.D. (Eds.) 1989. *Handbook of Research on Face Processing*. Elsevier Science Publishers B.V.

Yuille, A.L., Cohen, D.S., and Hallinan, P.W. 1989. Feature extraction from faces using deformable templates. In *Proc. Computer Vision and Pattern Recognition, CVPR-89*, pp. 104–109.

Yuille, A. and Hallinan, P. 1992. Deformable templates. In *Active Vision*, A. Blake and A. Yuille (Eds.), MIT Press: Cambridge, Mass, pp. 21–38.