Lip Posture Estimation using Kinematically Constrained Mixture Models

P. H. Kelly, E. A. Hunter, K. Kreutz–Delgado and R. Jain
Dept. of Electrical & Computer Engineering
University of California, San Diego

Abstract

A novel approach for estimating 3D lip posture from monocular video sequences is presented. The lips are modeled as a four body *closed kinematic chain* with each body possessing translational, rotational and prismatic (to account for deformations) degrees of freedom. Geometric constraints relating these bodies to each other, and to the face as a whole, are used to constrain the space of possible lip postures recovered from each image. These constraints are used with the recently proposed Expectation Constrained Maximization algorithm to estimate the lip posture from video frames that have been processed (using a color segmentation algorithm described here) to identify lip regions.

1 Introduction

This paper presents a new algorithm for estimating three–dimensional lip posture from video sequences. We model the lips as four interconnected bodies forming a *closed kinematic chain* in three–dimensional space. The bodies, or components, two for the upper, and two for the lower lip, possess translational, rotational and prismatic (i.e. sliding, to account for deformations) degrees of freedom. A set of model parameters determines the positions and orientations of these 3D components. These parameters are governed by a set of kinematic constraints enforcing the geometric structure of the lip components and their placement in the face. When imaged, lips in a face project to regions in the imaging plane. A color segmentation algorithm (described in Section 2) is used to identify these regions in video images. Our kinematic model is then used with the recently proposed Expectation Constrained Maximization (ECM) algorithm [11] to estimate the model parameter values using the segmented images. The use of kinematic constraints ensures that the recovered model parameters correspond to physically meaningful lip postures. Results demonstrate the efficacy of our approach indicating that model parameters, specifying three–dimensional lip position, can be estimated from realistic image data. Recovery of these lip parameters is important for many applications including visually assisted speech recognition (speechreading), expression recognition for intelligent interfaces, facial animation and low bit–rate video coding.

The computer vision literature is rich with techniques for facial analysis and synthesis. These include parametric flow models [3], optical flow with a finite element face model [6] and a finite element model for the lips [2]. Deformable contours, coupled with Kalman

BMVC 1998 doi:10.5244/C.12.8

filtering schemes [18, 12] have been employed to track facial movement in general and lips in particular. Detailed physically based facial models [19, 20, 16] have proven successful, especially for face synthesis activities. Point features are discussed in [13] in the context of face tracking and synthesis. A region based approach is adopted in [15] and used for tracking and analysis.

Like [15], we adopt regions as our underlying image "features." Regions (in our case, corresponding to lip segments) can be reliably extracted from images and provide robust evidence for object structure and posture. In our approach, however, we couple regions with a kinematic model of object structure. The use of a 3D model provides support against problems such as image noise and object occlusion, as well as avoiding issues such as image registration that must be addressed by appearance based methods.

The paper is organized as follows. Extraction of lip regions is outlined in Section 2 and the ECM algorithm is briefly reviewed in Section 3. Our kinematic lip model is then presented in Section 4, with results offered in Section 5. A summary concludes the paper.

2 Segmentation

Our algorithm requires segments corresponding to the mouth region along with some head orientation parameters. To provide a meaningful testbed, we have implemented both a color–based segmentation scheme and a simple correlation eye–tracker. (Eye locations provide enough head orientation information for use in our prototype development.) Together, these provide realistic raw measurements supporting adequate evaluation of our modeling approach.

We have developed a statistical, adaptive, color based segmentation scheme. Each pixel is characterized by a "feature" vector, namely, the average normalized color, $(\bar{nr}, \bar{ng})$ determined by averaging the normalized color, $\bar{nr} = \frac{R}{R+G+B}$ and $\bar{ng} = \frac{G}{R+G+B}$ (where R, G, B are the original color values at a pixel) in a small region around each pixel, e.g., a 5 by 5 neighborhood. Several classes, representing areas of interest in the face, are maintained. Each is characterized by a "feature" mean vector and covariance. Three such classes are employed, one for the lips and two for skin pixels accounting for the variability of skin color over the entire face. A user identifies areas of the head that correspond to these classes in the first frame of the sequence and the system determines the initial class statistics using this training data. Subsequent frames are automatically segmented as follows.

First, the head is segmented from the background using difference imaging between a background (i.e. object free) frame and the current frame. Next, a feature vector, i.e. the average normalized red and green values in a neighborhood of the pixel, is computed for each pixel in the input image. Each pixel is then classified as belonging to that class i such that $i = argmin_{i \in \mathcal{C}} (\mathbf{x} - \mu_\mathbf{i})^T \Sigma_i^{-1} (\mathbf{x} - \mu_\mathbf{i})$. Here, $\mathbf{x} = (\bar{nr}, \bar{ng})^T$ is the pixel's feature vector, $\mathcal{C} = \{1, 2, 3\}$ is the set of class indexes and Σ_i is the covariance matrix of the ith class.

Once all pixels have been classified, we perform a recursive update of the class statistics, akin to that performed in recursive least squares processing [10]. This update encourages the class statistics to track changes, such as those due to illumination, that cause the features to vary over time.

At each frame we thus produce a binary image whose "on" pixels correspond to those

pixels classified as belonging to the mouth region. These binary images are then smoothed using a symmetric Gaussian filter with small kernel and normalized so that the maximum value is 255. This filtering generates an image that can be interpreted as a scaled probability density giving, at each pixel, the likelihood that the pixel belongs to the mouth region.

Eye positions are recovered using a correlation tracker that is also manually initialized on the first frame. The tracker then determines eye locations in subsequent frames by selecting the point of highest correlation between a template (selected in the previous frame) and the new image. This has performed well on sequences where the head motion is small. More robust support to ensure adequate performance during events such as eye–blinks has not yet been implemented. Recent systems, e.g. [4] have demonstrated correlation based eye trackers that can handle such situations. These attributes could be easily incorporated into our current system.

Results from the segmentation and filtering process are shown in the figures in Section 5.

3 ECM Overview

The Expectation Constrained Maximization (ECM) algorithm [11] couples a probabilistic imaging model with knowledge of an articulated object structure (as modeled by open and closed kinematic chains) to recover articulated object posture from video sequences. The configuration in space of each body in the object is specified by the body's 3D mean (μ_i) and inertial Euler angle rotations (θ_i^1, θ_i^2) that are measured relative to a fixed coordinate system. These rotational parameters are used to form a rotation matrix, $\mathbf{R}_{\mathcal{O}i}$, giving the orientation of body i relative to an inertial frame $\mathcal{O}$. As each body is symmetric, only two rotational components are needed. Figure 1 schematically identifies these terms. All component 3D parameters are collected into the vector $\Theta \in \mathbf{R}^{5N}$ ($N = 4$ in our model), that completely determines the posture of the full multibody articulated object.

The ECM algorithm proceeds as follows. Each incoming frame is segmented and filtered as previously discussed. This smoothed image is interpreted as a probability distribution corresponding to a mixture of N Gaussians, one for the projection of each body onto the imaging plane. Thus, image sites, $\mathbf{x} = (x, y)$, are regarded as having been generated by a distribution with density, $\mathbf{p}(\mathbf{x}|\mathbf{\Phi}) = \sum_{i=1}^{N} \alpha_i p_i(\mathbf{x}|\phi_i)$. Mixture mode i is characterized by the mixing weight α_i (the *a priori* probability of component i), the statistics (mean and covariance, $\phi_i = \{\mu_{i,2D}, \Sigma_{i,2D}\}$) and the density, p_i. The assumption of Gaussian densities is not essential although it does make the mathematics fairly tractable. The lip geometry is shown in Figure 1; here $N = 4$.

Three–dimensional posture estimates for all components from the previous step (assumed valid), denoted $\Theta^{(s)}$, are projected, using scaled orthography[1] onto the segmented image. This projection provides an estimate of the two-dimensional mixture statistics, $\Phi^{(s)}$. The statistics are then updated via a step of the Expectation Maximization (EM) algorithm [5, 17], to produce component 2D statistics, Φ^{+}. The 3D inertial parameters are inferred (as discussed below) from these 2D statistics and collected into the vector

[1] A reasonable first order approximation to perspective projection [1]. An orthographic projection assumption has also been employed successfully in [19] to recover 3D face modeling parameters from dynamic contours tracked in an image.

Θ^+.

This intermediate 3D update is seldom kinematically feasible. Thus, we project this step back to a manifold determined by a set of non-linear constraints ($\mathbf{C}(\mathbf{\Phi}) = \mathbf{0}$) specifying kinematically feasible configurations. The update, Θ^+ is first projected to the constraint manifold tangent plane. This is followed by a Newton-Raphson (NR) iteration to bring this tangent plane projection onto the manifold proper [11, 8, 7]. At termination of the NR, we have determined $\Theta^{(s+1)}$, a new, kinematically valid, estimate of the model parameters. This estimate is then projected back onto the image plane, a new EM step taken and the process repeated. We iterate over each frame until the step generated by the EM is roughly orthogonal to the constraint manifold. This condition implies the model matches the data optimally in the least squares sense.

As noted above, the 3D parameters must be inferred from the 2D mixture statistics. Assuming scaled orthography, we can take the upper 2x2 matrix of the full 3D component covariance, $\Sigma_{i,3D}$ to be equivalent to the 2x2 observed 2D covariance matrix, $\Sigma_{i,2D}$ [9]. Equating terms, a system of non–linear equations can be formed relating the 3D inertial rotation parameters of each body (θ_i^1, θ_i^2) to the corresponding observed 2D statistics. To arrive at a closed form solution for the 3D orientations, additional knowledge in the form of shape and orientation constraints is employed. In [11] a shape constraint is enforced using a set of 3D model eigenvalues specifying the assumed shape (length and width) of each body in the chain. The observed 2D statistics are "clamped" to ensure they are consistent with these assumed 3D model eigenvalues. It can be shown that determination of the 3D θ_i^2 rotations can be done without explicit dependence on the numerical value of these 3D model eigenvalues. (The choice of these eigenvalues does exert an influence when the components are projected back into the image to begin another EM iteration.) The θ_i^1 rotations are computed in a similar fashion using an orientation constraint. The interested reader is referred to reference [11] for additional details on the ECM algorithm.

4 Kinematic Lip Model

This section develops the geometric constraint equations relating the lip components to each other and to the face as a whole. These equations determine the *constraint manifold* (the system $\mathbf{C}(\mathbf{\Phi}) = \mathbf{0}$) employed by the ECM algorithm described above in Section 3. These constraints ensure the kinematic structure of the lip model (e.g. specified interconnections between components) remains valid as we position the four body model at an appropriate location, orientation and extension so that it optimally aligns with the lip segment data.

The lips are located on the face and related to other facial features via several constraints outlined in this section. Figure 1 shows the overall geometry of the head and face and introduces notation that will be used below. The vector $\hat{n}_{Hs} = \frac{\mu_{e2} - \mu_{e1}}{\|\mu_{e2} - \mu_{e1}\|}$, gives the head sagittal plane normal. The head sagittal plane can be completely specified via this normal and the point, $p_{Hs} = \frac{\mu_{e2} + \mu_{e1}}{2}$, lying midway between the eyes.

The lips are represented using 4 interconnected ellipsoidal bodies, two for the upper lips and two for the lower. Figure 1 shows a two–dimensional projection of the 4 component lip model on a frontal face image. The links have fixed variance, or "size," but possess translational, rotational and prismatic (or sliding) degrees of freedom. As a result, they are free to rotate and slide along component centerlines (the $\hat{e}_{i1}$ axes). This

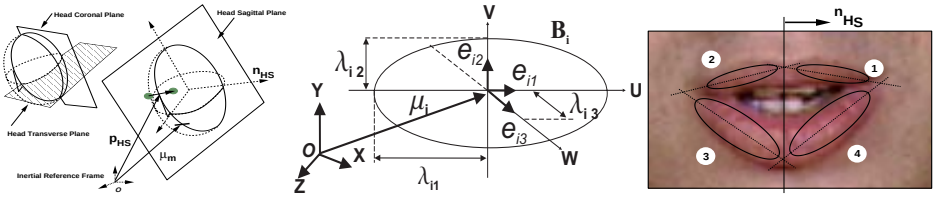

Figure 1: Head and mouth geometry and notation. Lip components have fixed variance (size, specified as $\lambda_{i1}, \lambda_{i2}$) but possess rotational and prismatic degrees of freedom accounting for variations in orientation and size. Body centerlines are specified by principal axes, $\hat{e}_{i1}$, determined from rotation matrix, $\mathbf{R}_{\mathcal{O}i}$ specifying body orientation with respect to inertial frame.

prismatic nature of each component accounts for the variability of lip size. While the components are free to slide along their centerlines, a variety of constraints, labeled *C1* to *C5* must hold and are discussed in the following paragraphs.

C1: General configuration of lip components. The four lip components are connected, thus intersections must occur between certain components. This constraint can be written as:

$$[(\hat{e}_{i1})(\hat{e}_{j1})\Delta\mu_{ij}] = 0 \tag{1}$$

Here, the tuples i, j assume the values $(1,2)$, $(1,4)$, $(2,3)$, and $(3,4)$. The notation, $[abc]$ indicates the scalar triple product ($[abc] = a \cdot (b \times c)$). Equation (1) forces the three vectors to be coplanar ensuring their intersection.

C2: Bounds on Component Intersection Distances The actual intersection distances between component i and j, denoted t_{ij} can be found in terms of the vectors $\hat{e}_{i1}$, $\hat{e}_{j1}$, and $\Delta\mu_{ij}$, according to the expression, $t_{ij} = \frac{[-(\Delta\mu_{ij} \times \hat{e}_{j1})\,(\hat{e}_{i1})(\hat{e}_{j1})]}{\|(\hat{e}_{j1}) \times (\hat{e}_{i1})\|^2}$. While not fixed, we do require these lengths to be bounded,

$$L_{ij}^2 \leq t_{ij}^2 \leq U_{ij}^2 \tag{2}$$

The values L_{ij} and U_{ij} are fixed *a priori* constants.

C3: Symmetry Constraints on Distance Terms. While equations (1) and (2) enforce a specific structure on the parameter estimation process, they may not be enough to ensure that a valid lip posture is recovered. Because the intersection distance parameters, t_{ij}, are not fixed, they may simply adjust to account for the data variations. A symmetry constraint on the distance terms acts to ameloriate this problem.

$$s_{ij}t_{ij} - s_{kl}t_{kl} = 0 \tag{3}$$

The tuples $\{(i,j);(k,l)\}$ take on the values $\{(1,2);(2,1)\}$, $\{(3,4);(4,3)\}$, $\{(1,4);(2,3)\}$, and $\{(4,1);(3,2)\}$, The terms, s_{ij}, take the value $+1$ or -1 to ensure the product, $s_{ij}t_{ij}$ is positive. (The intersection distances, t_{ij}, are signed.)

This constraint may limit the realizable lip posture space that our approach can recover. However, many lip motions, e.g. English speech, as well as common expressions, are fairly symmetric in nature. Also, it is unlikely that subtle asymmetries (often appropriate in animation) can be reliably recovered from the image regions we process. It is probably better, then, to estimate a reasonable structure reliably and modify the extracted data if appropriate, e.g. in the case where the estimates are used to drive an animation. Symmetry constraints also provide a mechanism for dealing with occlusion, e.g., as the head turns away from the camera so that far side lip components can no longer be observed. In this case, we can account for the unobserved components by reflecting the near side components parameters we can estimate (using the observed near side segments) about the head sagittal plane.

C4: Orientation Constraint. An orientation constraint is also employed to ensure the lip components wrap around the face in the gentle curve determined by the dental arc. The principal axes of the four lip components, taken two at a time, span a total of six planes. Two are of particular interest for development of this constraint. The first denoted, $\mathcal{P}_{14}$ is spanned by $\hat{e}_{11}$ and $\hat{e}_{41}$ with normal given by $\hat{n}_{14} = \frac{\hat{e}_{11} \times \hat{e}_{41}}{\|\hat{e}_{11} \times \hat{e}_{41}\|}$. Similarly, plane $\mathcal{P}_{23}$ is spanned by $\hat{e}_{21}$ and $\hat{e}_{31}$ and has normal, $\hat{n}_{23} = \frac{\hat{e}_{21} \times \hat{e}_{31}}{\|\hat{e}_{21} \times \hat{e}_{31}\|}$. We can constrain the orientation between these two planes using a variety of equations. We have found the following to be stable and effective.

$$\hat{n}_{ij} \cdot \hat{n}_{Hs} = \cos(\beta) \tag{4}$$

Here, (i, j) assumes the values $(1, 4)$ and $(2, 3)$. The angle β is a fixed constant for a particular subject. $\hat{n}_{Hs}$ is the head sagittal plane normal.

C5: Symmetry constraint about head sagittal plane. Finally, we force the intersection point between components 1 and 2, p_{12}, and that between components 3 and 4, p_{34}, to lie in the head sagittal plane. The following equation dictates the general condition,

$$\frac{p_{ij} - p_{Hs}}{\|p_{ij} - p_{Hs}\|} \cdot \hat{n}_{Hs} = 0 \tag{5}$$

Here, (i, j) assumes the values $(1, 2)$ and $(3, 4)$ specializing equation (5) for the two cases. p_{Hs} defines a point in the head sagittal plane midway between the eyes.

Results presented in the next section, do not reflect enforcing component/component intersection bounds. Enforcing these inequality constraints in the NR iteration is a current area of algorithm refinement.

5 Results

Data from three sequences is shown[2]. In the first, the subject is saying "Two, " in the second, saying the word "One," and in the third making a surprise expression. The first sequence was collected on a R5000 175 MHz SGI O2 at frame rate (30 fps). The second two sequences were collected using an R4600 132 MHz SGI Indy, at a sample rate of

[2]Video clips of our results are available on–line at `http://www-vision.ucsd.edu/~phkelly/Work/Work/face.html`.

about 3 frames per second. All three sequences were collected using a Sony Handy-Cam at 24-bit color and 320x240 resolution (further subsampled to around 160x120 pixels).

Results from the "Two" sequence are shown in Figure 2. The top row shows isodensity contours superimposed on the lip segments produced by our segmentation algorithm; this segment is the data our algorithm actually processes. The second row shows the same contours on top of the corresponding intensity image. Note, that the inner density contour does a good job of encircling the mouth interior region. The interior of this inner contour can be searched for the presence of the tongue and teeth, visible in some frames, which may prove useful for speechreading systems. As can be seen, the outer and inner isodensity contours track shape changes similar to active contours.

Spline fits to the posture parameters for the same frames are shown in the bottom two rows. Such curves are useful for tasks such as motion visualization and animation. In addition, such contours can provide useful features for analysis activities, such as visual speech recognition as powerfully demonstrated by Kaucic et al. [12]. The orthographic projection of the 3D spline onto the imaging plane is shown in the third row. The full 3D spline is shown at the bottom. The 3D spline is shown for each frame but from different view points. In the third 3D frame, note the curve corresponding to the bend of the lips around the face. This is a result of the orientation constraint *C4*. The generation of 3D data is an important distinction between our estimates and much dynamic contour work. The latter typically recovers estimates of contour location in the imaging plane. Our contour is an estimate of mouth location in 3D. Another, less significant difference, is that our contour, based as it is on the region data, lies roughly through the center of the lips rather than around the lips.

The splines curves were produced using cubic B–splines with control points determined from the estimated posture data. We take the control points equal to the component means and component/component intersections. The component means, as well as the horizontal intersections (i.e. between components 2 and 3 and between 1 and 4) are used as repeated control points so that the contour interpolates these points. This is consistent with our intuition that the component means are reliable estimates of lip position and should lie on a contour approximating the lips.

Figure 3 show selected frames from the "One" sequence and Figure 4 shows the "Surprise" sequence. An alternate visualization technique is used in these sequences. Namely, the projections of 3D body outlines and component axes onto the image. Because we are actually estimating a mixture, these component projections are, in effect, also isodensity contours. Indicating the posture in this manner is particularly useful to demonstrate the prismatic nature of the component/component joints that account for extension and compression of the lips. As can be seen, the algorithm effectively tracks changes in lip movement. It matches the filtered segments, the data it is actually tracking, quite well.

It is important to note that the "Surprise" sequence (Figure 4) consists of only the 4 frames shown (a result of the short expression duration and slow capture rate on the Indy). Thus, the difference in segment positions between the first and second frames is considerable, likewise for the third and fourth. Nevertheless, the algorithm remains stable and finds a reasonable posture at all frames in the sequence.

While our posture estimation system consistently recovers postures that match the segments, some discrepancies between the estimates and the intensity images are noticeable. This is primarily a result of the segmentation process that, as is typical of segmentation algorithms, will produce segments imperfectly matching the underlying object (in this

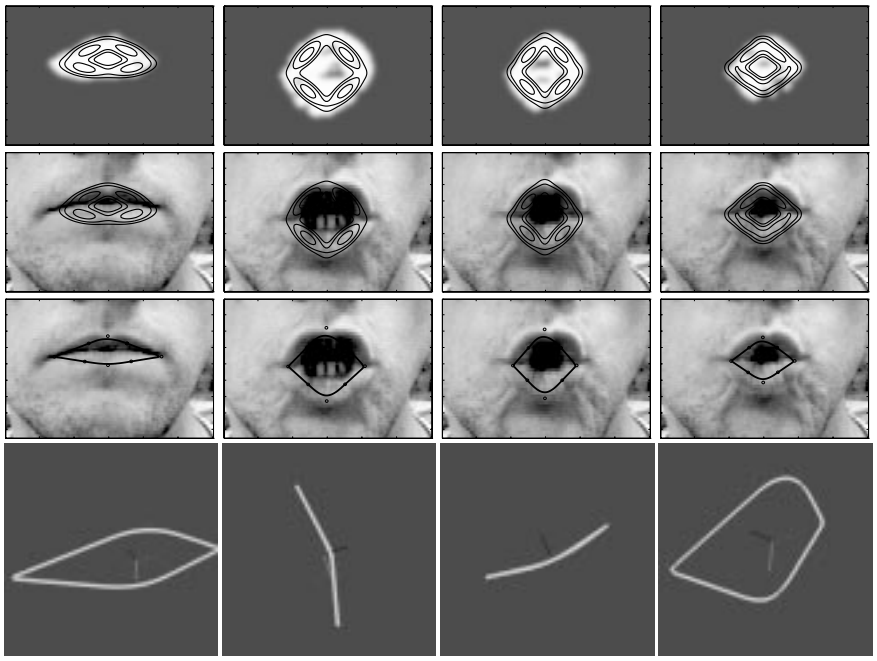

Figure 2: "Two" sequence. Top: isodensity contours superimposed on a segment. Second: contours superimposed on intensity data. Third: orthographic projection of 3D spline onto intensity image. Bottom: renderings of spline at each frame from different views. z-axis darkest and points into page, y-axis down. First frame: viewer looking toward right side of face. Second: towards left of face, in profile. Third: looking down on lips. Fourth: viewpoint similar to first.

case the lips). This is especially true of movements where the top lips pull up away from the teeth. For instance, in words such as "Test," expressions like a "toothy" smile, and in the central frames of the "Surprise sequence. In this case, very little segment is produced for the tails of the upper lip components and most of the component mass lies close to the top of the component near the intersection point. The estimation process thus produces components which tend to have excessive in–plane rotation, matching the segments, but not necessarily our more intuitive expectation of the mouth as defined by its corners.

The addition of more components to the model should ameloriate this problem. Towards this end we have designed and are currently evaluating a 6 body lip model that uses 2 modes for each of the upper lip components. Even without this modification, the current framework effectively captures the posture of the lips using realistic image segments.

6 Summary

This paper has delineated a novel, model–based lip representation and demonstrated its use for recovery of lip posture in a video sequences. The approach can serve as a substrate for the construction of facial analysis (e.g. expression recognition and speechreading) and

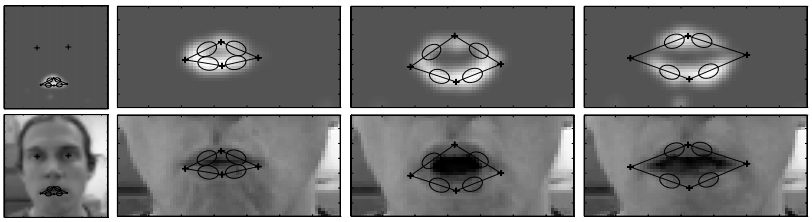

Figure 3: "One" sequence. Top: components projected onto segments. Bottom: components projected onto facial image. Note prismatic nature of the components. First frame shows neutral face with eye locations also identified for this sequence.

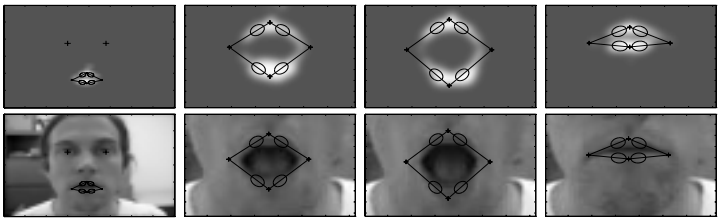

Figure 4: Several frames from the "Surprise" expression sequence.

synthesis (e.g. animation) systems. The representation describe herein is attractive for a variety of reasons. These include: *Use of regions:* parameters are estimated from image regions that can be reliably recovered from noisy image sequences using robust color segmentation techniques such as the scheme presented in Section 2. *Sparse image representation:* mouth posture is described by a small number (20) of parameters, considerably less than pixel based representations and fewer than some dynamic contour models. *Rich representation:* gross shape is determined from component means and intersection locations. Finer variations in mouth shape can be determined either directly from the estimated parameters as contours of constant density, or as a spline fit to the component means and intersection points. Other features can also be extracted from the model. For instance, we can compute the features mentioned by Massaro and Stork in their work on speechreading [14] (inner and outer width and height and horizontal separation of lip peaks). *3D model:* a three–dimensional model is employed and 3D parameters estimated. We believe this feature ensures consistent posture estimation despite large head motion, a hypothesis we are currently evaluating. Three–dimensional posture estimates may also prove useful for synthesis tasks.

References

[1] J. Y. Aloimonos. Perspective approximations. *Image and Vision Computing*, 8(3):179–192, August 1990.

[2] S. Basu and A. Pentland. A three–dimensional model of human lip motions trained from video. In *Proceedings IEEE Nonrigid and articulated motion workshop*, pages 46–53, San Juan, Puerto Rico, 17–19 June 1997. IEEE Computer Society.

[3] M. J. Black and Y. Yaccob. Recognizing facial expressions under rigid and non–rigid facial motions. In M. Bichsel, editor, *International workshop on automatic face and gesture recognition*, pages 12–17, Zurich, Switerland, June 26–28 1995.

[4] J. L. Crowley and F. Berard. Multi–modal tracking of faces for video communications. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, pages 640–645, San Juan, Puerto Rico, June 1997. IEEE Computer Society.

[5] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, 39(1):1–22, 1976.

[6] I. A. Essa, T. Darrell, and A. Pentland. Tracking facial motion. In *Proceedings of the Workshop on Motion of Non–Rigid and Articulated Objects*, pages 36–42, Austin, TX, November 11–12 1994. IEEE Computer Society Press.

[7] R. Fletcher. *Practical methods of optimzation*. Wiley, New York, 2nd edition, 1987.

[8] P. E. Gill, W. Murray, and M. H. Wright. *Practical Optimization*. Academic Press, New York, 1981.

[9] G. Grimmett and D. Stirzaker. *Probability and Random Processes*. Clarendon Press, Oxford, 1987. p. 69.

[10] S. Haykin. *Adaptive Filter Theory*, chapter 13. Prentice Hall, Englewood Cliffs, NJ, 2nd edition, 1991.

[11] E. A. Hunter, P. H. Kelly, and R. C. Jain. Estimation of articulated motion using kinematically constrained mixture densities. In *Proceedings IEEE nonrigid and articulated motion workshop*, pages 10–17, San Juan, Puerto Rico, 16 June 1997. IEEE Computer Society.

[12] R. Kaucic, B. Dalton, and A. Blake. Real–time lip tracking for audio–visual speech recognition applications. In *Proceedings European Conference Computer Vision*, Cambridge, UK, 1996.

[13] N. Magnenat-Thalmann, P. Kalra, and I. sunday Pandzic. Direct face–to–face communication between real and virtual humans. *International Journal of Information Technology*, 1(2):145–157, 1995.

[14] D. W. Massaro and D. G. Stork. Speech recognition and sensory integration. *American Scientist*, 86(3):236–244, 1998.

[15] N. Oliver, A. P. Pentland, and F. Berard. Lafter: lips and face real time tracker. In *Proceedings 1997 Conference on Computer Vision and Pattern Recognition*, pages 123–129, San Juan, Puerto Rico, 17–19 June 1997. IEEE Computer Society.

[16] F. I. Parke and K. Waters. *Computer Facial Animation*. A K Peters, Wellesley, Massachusetts, 1996.

[17] R. A. Redner and H. F. Walker. Mixture densities, maximum likelihood and the em algorithm. *SIAM Review*, 26(2):195–239, April 1984.

[18] D. Terzopoulos and R. Szeliski. Tracking with kalman snakes. In A. Blake and A. Yuille, editors, *Active Vision*, pages 3–20. Massachusetts Institue of Technology, 1992.

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

[20] K. Waters. A muscle model for animating three–dimensional facial expression. *Computer Graphics*, 21(4), 1987.