The Role of Middle Level Features for Robust Shape Tracking

Jacinto C. Nascimento[*] *Arnaldo J. Abrantes*[**] *Jorge S. Marques*[*]

ISR/IST ISEL ISR/IST

[*]Torre Norte, Piso 7, Instituto Superior Técnico, Av. Rovisco Pais, 1049-001 Lisboa, Portugal
[**] Rua Conselheiro Emídio Navarro, 1, 1949-014 Lisboa, Portugal

***Abstract* – Shape trackers using low level features (e.g., edge points) often fail in complex environments (e.g. clutter, inner edges or multi-objects). Two alternatives are discussed in this paper. Both methods use middle level features: (data centroids, strokes), which are more informative and reliable than edge transitions used in most shape trackers. Furthermore, it is assumed that each middle level feature is either valid or an outlier. Therefore a confidence degree is assigned to each feature. Features with a high degree of confidence have a large influence on the shape estimate while features with low degree of confidence have a negligeable influence on the final estimates. Both mechanisms (the use of middle level features and confidence degree) lead to a significant improvement of the tracker robustness. This is shown in the paper in the context of lip tracking problem.**

***Keywords* – Middle level features, shape analysis, deformable models, robust methods.**

I. Introduction

Object tracking is a challenging problem which has been extensively studied for more than two decades [1]. Active contours are among the most popular approaches. They represent the object boundary by an elastic model attracted by low level features, e.g., edge points [2] or intensity transitions obtained by directional search [3]. The estimation of the model parameters is often performed by Kalman or particle filtering [4], [5]. Unfortunately, feature detection is an ill-posed problem and the low level features detected in the image often contain outliers which limit the performance of shape tracking algorithms (see figure 1).

Several methods were proposed to alleviate this difficulty. Some impose additional restrictions to the object shape by adopting rigid templates [6], or eigen shapes learned from the data [7]. Temporal restrictions have also been considered by describing the evolution of the motion and shape parameters using stochastic difference equations. The parameters of the dynamic model can also be trained from video sequences using standard identification methods [8], [9].

None of these approaches however is enough to solve the segmentation problem i.e., to discriminate valid data from the outliers which hamper the performance of the shape trackers. An *ad hoc* procedure used to improve Kalman estimates consists of using a validation gate computed from the predicted object boundary, discarding the observations which are outside [7]. This method works well if the object motion is slow and predictable but fails in more complex situations. Nonlinear filtering methods based on non-Gaussian noise distributions have also been used to address this problem [5].

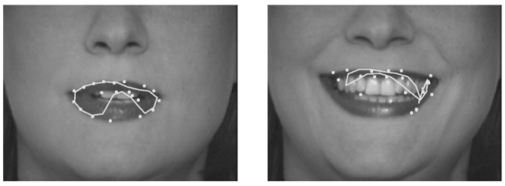


Figure 1 - Shape estimation with outliers.

This paper considers two robust filtering methods: a centroid based method and a method denoted as Shape Probabilistic Data Association Filter (S-PDAF). The first algorithm performs a fuzzy labeling of low level features (edge points) detected in the image, considering each of them either as reliable or outlier. A degree of confidence is assigned to each feature using an object model and a noise model. The object model is a mixture of Gaussians which attempts to represent the boundary points of the object in each new frame while the outliers are represented by a single Gaussian distribution with a large covariance. The mixture parameters are interpreted as middle level features which summarize the image data useful for the estimation of the object boundary. This approach is related to the work presented in [10], [11] in which intermediated representations of the data are used for compression and robustness purposes.

The second algorithm organizes the low level features (edge points) in strokes. Instead of assigning a degree of confidence to each stroke, the S-PDAF considers sequences of stroke labels and computes a degree of confidence for each sequence. Each sequence corresponds to a classification of all the strokes detected in the image as valid or outlier. The update of the model parameters (B-spline coefficients) is inspired in the probabilistic data association filter (PDAF) proposed by Bar-Shalom and Fortmann in the context of target tracking in clutter [12], providing a fast and robust filtering algorithm.

This work was partially supported by FCT in the scope of TMO project. Corresponding author: Jacinto Nascimento, E-mail: `pcjan@pop.isr.ist.utl.pt`

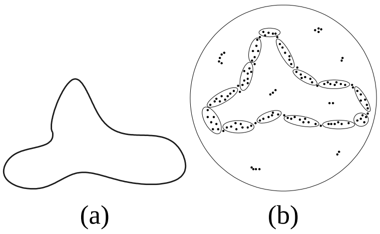


Figure 2 - (a) Object boundary, (b) low level features and mixture model (middle level features).

Results in traffic, lip tracking experiments are provided to evaluate both algorithms. As expected, both methods perform better than the classic Kalman trackers.

II. Centroid Based Method

Let I denote an image and $y = \{y_1, \ldots, y_N\}$ the coordinates of the edges points detected in I. We wish to approximate a subset of y by a parametric curve $c_t(s)$, where s is the arc length.

Two difficulties have to be considered: first we need to know which edge points are valid (belong to the object contour) and which are outliers. Second we need to match each valid feature with a contour point (matching problem).

To overcome these difficulties it will be assumed that the edge points y_i are independent random variables. Their distribution is described by a mixture of $N+1$ Gaussians:

$$p(y_i) = \sum_{k=1}^{N+1} \alpha_k \mathcal{N}(y_i; \mu_k, R_k) \qquad (1)$$

where $\mathcal{N}(y_i; \mu, R)$ denotes a normal distribution with mean μ and covariance matrix R and α_k are the mixture coefficients. The first N Gaussians are used to represent the object contour and their means are located at equally spaced points on the curve $c_t(s)$. The last Gaussian has a large covariance matrix and it is used to represent the invalid data points (outliers), see figure 2.

To solve the above mentioned difficulties (data segmentation and matching) all we need is to estimate which Gaussian generated each edge point y_i. Since it is not possible to obtain an error free classification of the edge points, a set of probabilities (confidence degrees) will be assigned to each feature. The estimation of the mixture parameters and confidence degrees is recursively performed using the Expectation-Maximization (EM) algorithm [13].

The previous approach is valid when a single image is available. In tracking problems we have to merge past information about the object shape and uncertainty with the current data. This will be done using Kalman filtering. It will be assumed that the object boundary c depends on a set of parameters $x_t \in \Re^n$ which $c_t(s) = c(s, x_t)$ evolves in time. The evolution of x_t is usually described by a discrete state model with random input

$$x_t = Ax_{t-1} + w_t \qquad (2)$$

where $w_t \sim \mathcal{N}(0, Q)$. In the proposed algorithm, the Kalman filter is not driven by low level features directly, but with the mixture parameters instead. The means of the first N Gaussians are considered as noisy observations of the object boundary at N sample points

$$\mu_i = c(s_i, x_t) + v_i \qquad (3)$$

where v_i is an observation noise with zero mean and covariance R_i estimated before in the EM step.

III. S-PDAF

The S-PDAF is a robust tracking algorithm recently proposed in [14]. The main ideas are simple. First, the low level features (intensity transitions obtained by directional search along lines orthogonal to the object boundary) are organized in M strokes. Two neighboring features are linked when their distance to the predicted contour is similar. It is assumed that some strokes belong to the boundary of the object to be tracked (valid strokes) while the others do not (outliers). Since each stroke may be valid or invalid, 2^M hypotheses (interpretations) have to be considered.

The object shape and position are described as in the previous section, however, the output equation depends on the interpretation i.e., we have a bank of output equations each one associated to a different data interpretation

$$y_{it} = C_{it} x_t + \eta_t \qquad (4)$$

where C_i is the observation matrix associated to the i-th interpretation and $\eta_t \sim \mathcal{N}(0, R_i)$ is a white Gaussian measurement noise. The observation matrices C_i, C_j associated with two interpretations I_i, I_j are always different since they represent different data points (outliers are not described by (4)).

Since there are multiple interpretations the propagation of the *a posteriori* density of the state vector $p(x_t \mid Y^t)$ given the current and the past observations Y^t, is a mixture of Gaussians, with a number of modes increasing with t [15]. To avoid the combinatorial explosion associated with the use of multiple interpretations, it will be assumed that the distribution of the unknown parameters x_t, given past observations Y^{t-1}, is Gaussian, i.e.,

$$p[x_t \mid Y^{t-1}] = \mathcal{N}[x_t; \hat{x}_t^-, P_t^-] \qquad (5)$$

where $\hat{x}_t^-$, P_t^- are the mean and covariance of x_t given Y^{t-1}.

The computation of the state estimate and uncertainty given the current and the past observation are given by,

$$\hat{x}_t \triangleq E[x_t \mid Y^t] \qquad (6)$$

$$P_t \triangleq E\left\{[x_t - \hat{x}_t][x_t - \hat{x}_t]^T \mid Y^t\right\} \qquad (7)$$

Since it is possible to have multiple data interpretations, they all have to be considered. Equation (6) can be written as

$$\hat{x}_t = \sum_i \alpha_{it} \int x_t p(x_t \mid I_{it}, Y^t)\, dx_t \qquad (8)$$

Equation (8) can be rewritten as

$$\hat{x}_t = \sum_i \alpha_{it} \hat{x}_{it} \tag{9}$$

where $\alpha_{it} \triangleq p(I_{it} \mid Y^t)$ is the *a posteriori* probability of the i-th interpretation (data association probability), and

$$\hat{x}_{it} = E\{x_t \mid I_{it}, Y^t\} \tag{10}$$

The state estimate $\hat{x}_t$ is a weighted sum of the state estimates $\hat{x}_{it}$ obtained for each interpretation I_{it} and updated by Kalman filtering

$$\hat{x}_{it} = \hat{x}_t^- + K_{it} \nu_{it} \tag{11}$$

where $K_{it}, \nu_{it} = y_{it} - C_i \hat{x}^-$ are the Kalman gain and innovation for the interpretation I_{it}. Replacing (11) in (9)

$$\hat{x}_t = \hat{x}_t^- + \sum_{i=1}^{m_k} \alpha_{it} K_{it} \nu_{it} \tag{12}$$

A recursive equation can also be derived for the covariance matrix (see [14]).

$$P_t = \left[I - \sum_{i=1}^{m_k} \alpha_{it} K_{it} C_i \right] P_t^- + \sum_{i=0}^{m_k} \alpha_{it} \hat{x}_{it} \hat{x}_{it}^T - \hat{x}_t \hat{x}_t^T \tag{13}$$

Equations (12, 13), define the state and uncertainty update for S-PDAF.

Since we do not know which interpretation is valid, a probability α_{it} (confidence degree) is computed for each data interpretation. This requires a probabilistic model for the image strokes. Three items will be considered in this model: the stroke lengths, the distances from the stroke points to the best estimate of the object boundary and stroke superposition. The following variables are used to describe the association probability: M - number of strokes detected in the image; b,e - vectors with the first and last indices of all data strokes; I_i - interpretation. The *a posteriori* probability of the i-th interpretation is

$$\alpha_{it} = P\{I_{it} \mid y_t, b, e, M, Y^{t-1}\} \tag{14}$$

which can be decomposed as follows

$$\alpha_{it} = c\, p(y_t \mid I_{it}, b, e, M, Y^{t-1})\, p(I_{it} \mid b, e, M, Y^{t-1}) \tag{15}$$

where y_t denotes the data observed at instant t and c is a normalization constant.

The first density function in (15) depends on the distance between the strokes to the predicted contour. Assuming that all features are independently generated, it is assumed that this probability is the product of normal distributions associated to the features considered as reliable in the i-th interpretation and a product of uniform distribution for the unreliable ones. The second density in (15) assumes that the stroke labels are independent and long strokes have higher probability than short ones. This can be accomplished by using a linear model to represent the dependence of stroke probability with length (details can be found in [14]).

Figure 3 shows a simple example with two strokes. The association probabilities are displayed in Table I. In this example, the most probable interpretation ($\alpha = 0.5006$) is the one which considers both strokes as valid. Finally, the a posteriori distribution of the unknown parameters is computed, considering all possible interpretations weighted by their associated probabilities.

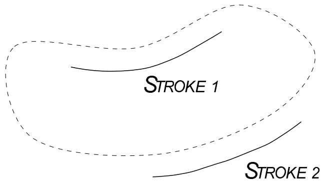


Figure 3 - Feature detection: predicted contour (dashed line) and detected strokes (solid line)

S1	*S2*	α
0	0	0.0852
0	1	0.1915
1	0	0.2226
1	1	0.5006

Table I

DATA INTERPRETATIONS AND ASSOCIATION PROBABILITIES

IV. SHAPE MODEL

To represent a moving object in a given frame t, it is assumed that the object boundary is a transformed version of a reference shape plus an additional deformation. Let $c_t(s) : I \rightarrow \Re^2$ be a parametric representation of the reference curve, where $I \in \Re$, $s \in I$ is a parameter defining the location of a point in the curve, and $c^r(s)$ a reference shape of the object (which can be obtained in the first image of the sequence). It is assumed that

$$c_t(s) = \mathcal{T} c^r(s) + c_t^d(s) + \eta_t(s) \tag{16}$$

where $\eta_t(s)$ is a noise curve and $\mathcal{T}$ is a geometric transformation, $c_t(s)$, $c^r(s)$, $c_t^d(s)$, are parametric descriptions of the object shape, reference shape and deformation respectively. In this paper B-splines are used.

Several transforms can be considered (e.g., translation, Euclidean similarities, affine transform). For instance in the case of affine transformation equation (16) can rewritten as

$$\left\{ \begin{array}{l} c_{1t}(s) = x_{1t} c_1^r(s) + x_{2t} c_2^r(s) + x_{3t} + c_{1t}^d(s) + \eta_{1t}(s) \\ \\ c_{2t}(s) = x_{4t} c_1^r(s) + x_{5t} c_2^r(s) + x_{6t} + c_{2t}^d(s) + \eta_{2t}(s) \end{array} \right. \tag{17}$$

where $c_t(s) = (c_{1t}(s), c_{2t}(s))$, $c^r(s) = (c_1^r(s), c_2^r(s))$, $c_t^d(s) = (c_{1t}^d(s), c_{2t}^d(s))$, $\eta_t(s) = (\eta_{1t}(s), \eta_{2t}(s))$, and $x_{1t}, \ldots, x_{6t}$ are unknown parameters to be estimated.

Dynamic equations must be considered to represent the evolution of the model parameters. Let x_t denote the vector of unknown shape, motion and deformation parameters

$$x_t = [x_{1t}, \dots, x_{Dt}, \dot{x}_{1t}, \dots \dot{x}_{Dt}, d_{11}, \dots, d_{1L}, d_{21}, \dots, d_{2L}]^T \tag{18}$$

where D is the dimension of the shape space, and d_{iL} are the deformation parameters at the L control points.

Let y be a $2N$ vector obtained by sampling the object at N equally spaced points

$$y = [c_{1t}(s_1), \dots, c_{1t}(s_N), c_{2t}(s_1), \dots, c_{2t}(s_N)]^T \tag{19}$$

Equation (17) can be rewritten as in (3), (4) with

$$C = \begin{bmatrix} M & \mathbf{O}_{N\times 3} & \mathbf{O}_{N\times 6} & \mathbf{B}_{N\times L} & \mathbf{O}_{N\times L} \\ \mathbf{O}_{N\times 3} & M & \mathbf{O}_{N\times 6} & \mathbf{O}_{N\times L} & \mathbf{B}_{N\times L} \end{bmatrix} \tag{20}$$

where

$$M = \begin{bmatrix} c_1^r(s_1) & c_2^r(s_1) & 1 \\ c_1^r(s_2) & c_2^r(s_2) & 1 \\ \vdots & \vdots & \vdots \\ c_1^r(s_N) & c_2^r(s_N) & 1 \end{bmatrix} \tag{21}$$

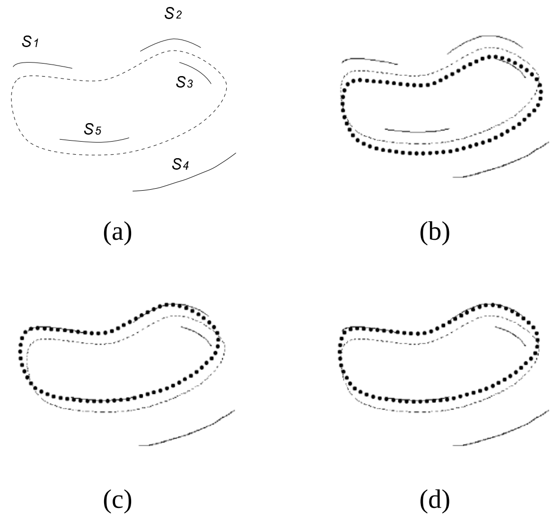
and B is the interpolation B-spline matrix, $\mathbf{O}$ is the null matrix. Similar expressions can be obtained for the other transformations. A tutorial about shape space models can be found in [4].

V. Results

Experimental tests were performed to evaluate the three algorithms: the Kalman tracker, the centroid based tracker and the S-PDAF. Synthetic and real images were used to assess the performance of these methods. It was observed that in these experiments the centroid based method is 3.7 slower than the Kalman tracker, and the the S-PDAF is 3 times slower than the Kalman tracker.

Figure 4 shows a synthetic example. The dashed line represents the initial (predicted) contour, while the solid lines represent the detected strokes in the image. The goal is to estimate the object boundary which best matches to the strokes localization. Looking at figure 4 (a), it is easily concluded that the best interpretation classifies S_1, S_2, S_5 as valid data and S_3, S_4 as outliers, assuming that the object undergoes a translation motion (2 degrees of freedom). The results obtained by the three algorithms using a translation model without deformation are shown in figure 4 (b)-(d). The shape estimates obtained with the robust methods (c,d) described in this paper are much better than the estimates obtained with the Kalman tracker (b) since they manage to neglect the influence of the outlier strokes and provide correct estimates of the translation using S_1, S_2 and S_5. This happens since these two methods use higher level features and robust estimation methods.

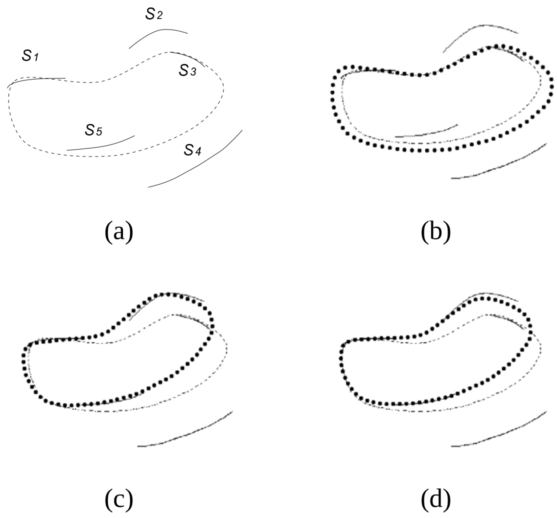
Figure 5 illustrates another situation. Here the position and orientation suggests that the object boundary undergoes a translation and rotation motion. Euclidean similarities are used to model the motion of the object. Thus 4 degrees of freedom are used in the state vector, allowing translation and rotation. The centroid based tracker and the S-PDAF solve this problem well, estimating the rotation and discarding the influence of the outliers strokes. However, the Kalman filter gives a wrong answer. The object boundary is attracted towards the outlier stroke S_4.



(a) (b)

(c) (d)

Figure 4 - (a) Initialization, results obtained with: (b) Kalman filter, (c) Centroid based tracker, (d) S-PDAF.



(a) (b)

(c) (d)

Figure 5 - (a) Initialization, and results obtained with: (b) Kalman filter, (c) centroid based tracker, (d) S-PDAF.

Figure 6 show the tracking performance of each algorithm in traffic sequences. The motion model used in this examples was the same as in the previous example ($x \in \Re^4$) allowing shape translations, rotations and scaling on the car boundary. It is concluded that the Kalman filter, figure 6 (left column), provides wrong estimates of the rotation angle and translation vector. The shape estimates is being attracted towards the inner edges of the car to be tracked. This behaviour was already observed before in figure 5. The results obtained by the robust trackers are much better. Once again the confidence degrees assigned to the features play a crucial role to achieve robust estimates of the object boundary.

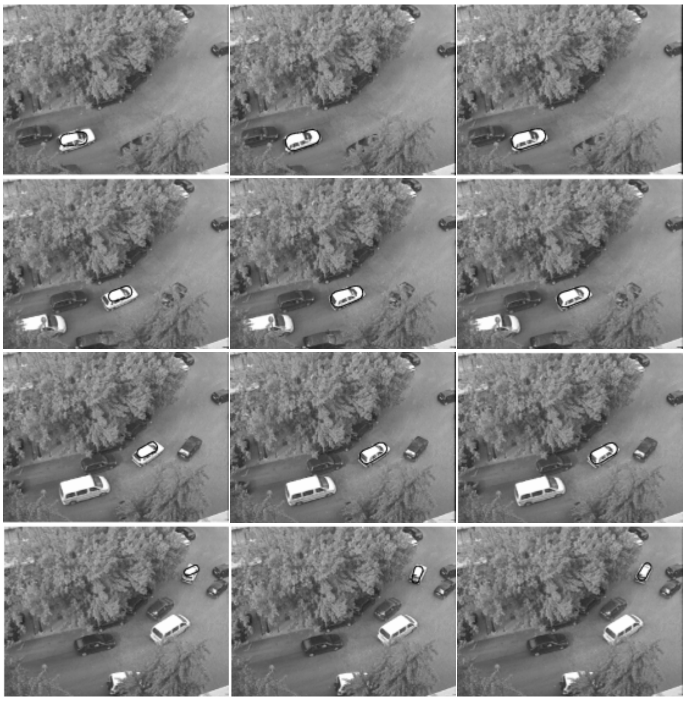


Figure 6 - Tracking results with Kalman (left), Centroid based (center) and S-PDAF (right) (frames 8, 21, 32, 57), (dark line is the output of the algorithms).

Figures 7, 8 and 9 display the performance of the algorithms in lip tracking. In these experiments an affine map with deformation in control points is used. These experiments illustrate the estimation of the lips boundaries while a person is talking (figures 7,9) and singing (figure 8). In the examples of figures 8,9 the lip boundary suffer sudden changes while in figure 7 the motion is smoother. In figures 7 and 8 the dots (in Kalman and S-PDAF trackers) and the thin white lines (in the centroid tracker) are the detected feature. In figure 9 the observations are not displayed for the sake of visualization. The centroid based tracker and the S-PDAF are able to track the lips during the whole sequence while the Kalman tracker becomes lost after the first 25 frames (see figure 7).

In singing example the Kalman tracker loses the lips boundaries as before (see figure 8 (left)). The centroid based tracker tends to lose track during short intervals specially in the presence of abrupt motion changes but manages to recover after a few frames. The same happens in the final example (figure 9).

VI. Conclusions

Two tracking algorithms are presented and tested in this paper using middle level features, i.e., features which are more informative than edges or corners. Both algorithms assume that the features detected in the image are not reliable. Some of them are outliers. A degree of confidence is computed for each feature. Therefore, the contribution of each feature for the estimation of the object boundary depends on the degree of confidence.

While acceptable results were achieved by both methods in experiments in which the Kalman tracker fails, the best results were obtained by the S-PDAF tracker which exhibits a remarkable ability to simultaneously cope with outliers and abrupt shape changes.

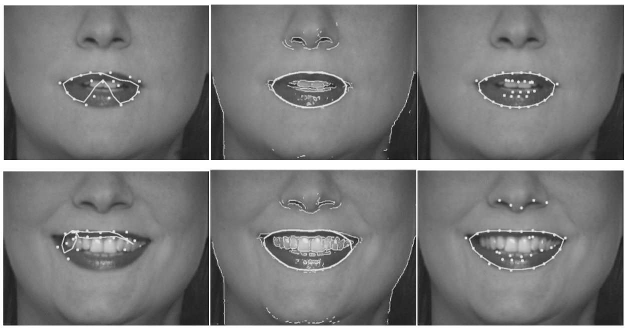


Figure 7 - Tracking results with Kalman (left), centroid based (center) and S-PDAF (right) (frames 28,38). White points represent low level features.

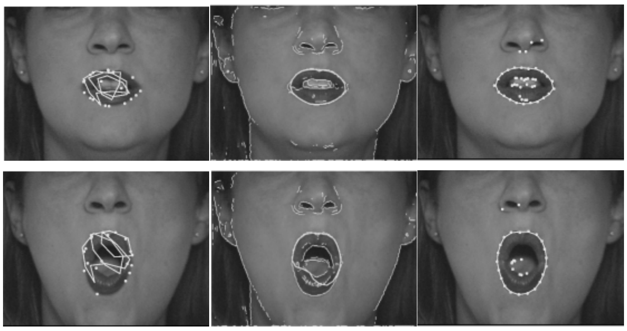


Figure 8 - Tracking of abrupt transitions with Kalman (left) centroid based (center) and S-PDAF (right) (frames 31,32). White points represent low level features.

References

[1] H. Nagel, "Image sequence evaluation: 30 years and still going strong", *ICPR*, vol. 1, pp. 149–158, 2000.

[2] L. Cohen and I. Cohen, "Finite element methods for active contour models and ballons for 2-d and 3-d images", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 15, no. 11, pp. 1131–1147, 1993.

[3] H. Tagare, "Deformable 2d template matching using orthogonal curves", *Trans. Medical Imaging*, pp. 108–117, 1997.

[4] A. Blake and M. Isard, *Active Contours*, Springer, 1998.

[5] M. Isard and A. Blake, "A mixed-state condensation tracker with automatic model-switching", in *Int. Conference on Computer Vision*, pp. 107–112. 1998.

[6] A. Blake, R. Curwen, and A. Zisserman, "A framework for spatiotemporal control in the tracking of visual contours", *Int. Journal of Computer Vision*, vol. 2, no. 11, pp. 127–145, 1993.

[7] T. Cootes, T. Hill, C. Taylor, and J. Haslam, "The use of active shape models for locating structures in medical images", *Image Vision and Computing*, vol. 12, pp. 355–366, 1994.

[8] A. Blake, M. Isard, and D. Reynard, "Learning to track the visual motion of contours", *Artificial Intelligence*, vol. 78, pp. 179–212, 1995.

[9] A. Baumberg and D. Hogg, "Learning deformable models for tracking the human body", in *Motion-Based Recognition*, R. Jain M. Shah, Ed., pp. 39–60. Kluwer, 1997.

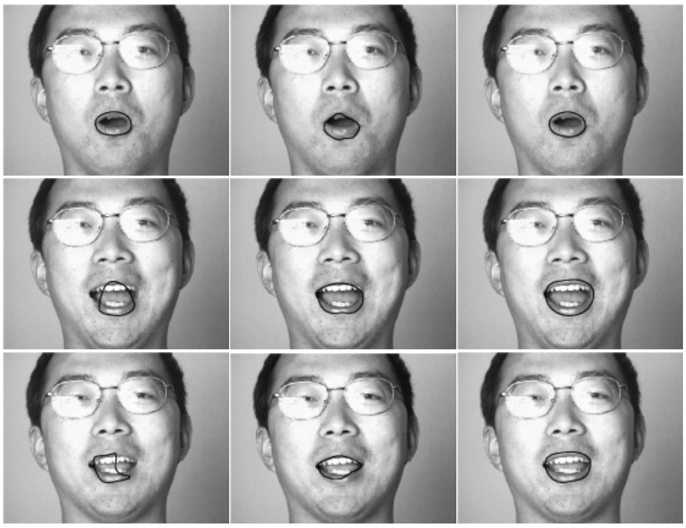


Figure 9 - Tracking results with Kalman (left) centroid based (center) and S-PDAF (right) (frames 11, 16,67).

[10] A. J. Abrantes and J. S. Marques, "A class of constrained clustering algorithms for object boundary detection", *IEEE Trans. on Image Process*, vol. 5, no. 11, pp. 1507–1521, 1996.

[11] Y. Wu, Q. Tian, and T. Huang, "Integrating unlabeled images for image retrieval based on relevance feedback", *ICPR*, vol. 1, pp. 21–24, 2000.

[12] Y. Bar-Shalom and T. Fortmann, *Tracking and Data Association*, Academic Press, 1988.

[13] A. Dempster, M. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the em-algorithm", *Journal of the Royal Statistical Society*, vol. B(39), pp. 1–38, 1977.

[14] J. Nascimento and J. Marques, "Robust shape tracking in the presence of cluttered background", *Int. Conf. on Image Processing*, pp. 82–85, September 2000.

[15] J. Tugnait, "Detection and estimation for abruptly changing systems", *Automatica*, vol. 18, pp. 607–615, 1982.