Automatic Facial Landmark Tracking in Video Sequences using Kalman Filter Assisted Active Shape Models

Utsav Prabhu, Keshav Seshadri, and Marios Savvides
uprabhu@andrew.cmu.edu, kseshadr@andrew.cmu.edu, msavvid@ri.cmu.edu

Department of Electrical and Computer Engineering, Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, PA 15213, USA
<http://www.ece.cmu.edu>

Abstract. In this paper we address the problem of automatically locating the facial landmarks of a single person across frames of a video sequence. We propose two methods that utilize Kalman filter based approaches to assist an Active Shape Model (ASM) in achieving this goal. The use of Kalman filtering not only aids in better initialization of the ASM by predicting landmark locations in the next frame but also helps in refining its search results and hence in producing improved fitting accuracy. We evaluate our tracking methods on frames from three video sequences and quantitatively demonstrate their reliability and accuracy.

1 Introduction

Active Shape Models (ASMs) aim at automatically locating key landmark points that define the shape of any object. The technique has been used to a great extent in medical applications [1], [2] and particularly in locating facial features [3], [4], [5], [6]. A primary requirement in the ASM fitting process is that the object's position in the image must be known in order to initialize it and for facial landmarking this necessitates the use of a face detector.

This paper is concerned with the problem of automatically locating facial landmarks of a single person in successive frames of a video sequence using an ASM. Knowing the location of such facial features can aid in face recognition [7], expression analysis [8] and pose estimation [9] and it is for these reasons that there has been a lot of prior work in this field over the past few years. When dealing with a video that does not exhibit scene change (such as surveillance footage), an approach that utilizes information about landmark positions in the previous frame to initialize the ASM for the current frame is expected to produce better results than a method that treats each frame independently. Such an approach is easy to implement and certainly avoids the errors associated with face detection across multiple frames, but relies on the ASM fitting being extremely accurate. Since this is not always the case, the need for refining the results obtained by an ASM becomes clear. A discrete Kalman filter [10] can be used for this purpose and by treating the landmark coordinates that are

fitted by an ASM as noisy measurements, it can correct them to produce results which are much closer to the true locations of the landmarks. Its motion model also allows the prediction of landmark positions in the next frame using their positions in the previous frame. Thus, Kalman filtering serves in providing good initialization and in correcting ASM results.

We propose two implementations that utilize Kalman filtering to track landmarks in different ways. We evaluate these implementations on frames from three video sequences and by comparing the results to the coordinates of landmarks that were manually annotated, are able to quantitatively assess their goodness of fit. Both methods show good tracking ability even when dealing with large in-plane rotation and pose variation of faces which is significant given that the ASM we use is trained only on frontal faces.

The rest of this paper is organized as follows. Section 2 provides an overview of the theory behind Kalman filters and ASMs as well as a short description of the particular ASM implementation we use for landmark tracking in this paper. Section 3 describes related work that has been carried out on landmark tracking in videos. In section 4 we propose and describe the various schemes for landmark tracking that we will be evaluating while section 5 describes the results that were obtained using them. Finally, section 6 presents some conclusions on this work and addresses possible areas of improvement.

2 Background

2.1 Active Shape Models (ASMs)

An ASM builds a point distribution model and a local texture model of the region around pre-defined points (referred to as landmarks) placed along the contours of an object. For applications in automatic facial landmarking, an ASM must first be trained on images in which facial landmarks have been labeled by hand. A set of N landmarks placed along the contours of a face constitute a shape vector $\mathbf{x}$ in which the x and y coordinates of each landmark are stacked as shown in (1).

$$\mathbf{x} = (x^1, x^2, \dots, x^N, y^1, y^2, \dots, y^N)^T \quad (1)$$

Once the shape vectors for each of the n faces in the training set have been obtained, they are aligned using Generalized Procrustes Analysis [11] and the mean shape $\bar{\mathbf{x}}$ and covariance matrix $\mathbf{S}$ of these aligned shapes are calculated.

Principal Component Analysis (PCA) is now applied on the shape vectors by computing the eigenvalues and eigenvectors of $\mathbf{S}$. The eigenvectors corresponding to the eigenvalues that model most (95-98%) of the variation are retained along the columns of a matrix $\mathbf{P}$. It is now possible to represent the positions of landmarks in any face by the shape model equation given in (2)

$$\mathbf{x} \approx \bar{\mathbf{x}} + \mathbf{P}\mathbf{b} \quad (2)$$

where $\mathbf{b}$ is a vector of the PCA coefficients of the shape $\mathbf{x}$.

Fig. 1. Landmarking scheme used by the Modified ASM (MASM) implementation.

Statistics of the local texture around each landmark are obtained by constructing 1D or 2D profiles. 1D profiles sample the gray level pixel intensities along lines through a landmark and perpendicular to the shape boundary while 2D profiles sample these intensities in square regions around a landmark. Old implementations such as those described in [3], [12], [13] used 1D profiles but more recent implementations such as those in [4], [5] use 2D profiles which capture more information and hence contribute to better fitting results. Regardless of whether 1D or 2D profiles are used, the mean profile $\bar{\mathbf{g}}$ and covariance matrix $\mathbf{S}_{\mathbf{g}}$ of all profiles in the training set are computed for multiple resolutions (typically 3-4) of an image pyramid.

Once the shape model and local texture model have been built, the ASM can be used to locate landmarks in a test image. A face detector provides the initialization needed for this purpose. The mean shape is aligned over the face detected and profiles are constructed for several locations in the neighborhood of each landmark. A landmark is moved to the location that has a profile $\mathbf{g}$ with the lowest Mahalanobis distance to the mean profile $\bar{\mathbf{g}}$ for the landmark in question as given by $f(\mathbf{g})$ in (3).

$$f(\mathbf{g}) = (\mathbf{g} - \bar{\mathbf{g}})^T \mathbf{S}_{\mathbf{g}}^{-1} (\mathbf{g} - \bar{\mathbf{g}}) \quad (3)$$

The process is repeated until the landmark locations do not change by much between iterations or a certain number of iterations is exceeded. After each iteration, the PCA shape coefficients are obtained and are constrained by clipping their values (b_i) to lie in the range $-3\sqrt{\lambda_i}$ to $+3\sqrt{\lambda_i}$ where λ_i is the eigenvalue corresponding to the i^{th} eigenvector retained by PCA. This process ensures that the fitted shapes are representative of the training shapes. Once convergence is

reached at the lowest resolution image, the landmark positions are scaled and utilized as the starting shape for an image of higher resolution in the image pyramid and the final positions are available once convergence is reached at the highest resolution image.

The ASM implementation we use for landmark tracking in this paper is the Modified ASM (MASM) approach described in [5], which has been shown to automatically landmark faces with a fair degree of accuracy. This implementation uses 79 landmarks to model the contours of a face (shown in Fig. 1) and the OpenCV 1.0 implementation [14] of the Viola-Jones face detector [15] for initialization of the model. The images that are used to train MASM are a set of 4000 manually annotated images from the query set of the still face challenge problem of the National Institute of Standards and Technology (NIST) Multiple Biometric Grand Challenge - 2008 (MBGC-2008) [16], [17] which consists of 10,687 images of 570 subjects.

2.2 The Discrete Kalman Filter

The Kalman filter has proved to be an essential tool for real-time signal tracking, with wide-spread applications in control systems, navigation and computer vision. Designed to function as a predictive-corrective algorithm, it aims at finding the optimal estimate of state in a linear dynamic system. This state estimate is a weighted average of the predicted system state and a noisy measurement. The weights are calculated from the state covariance at each time step, and represent the trust in the system prediction versus the measurements.

The discrete Kalman filter addresses the general problem of trying to estimate the state $\mathbf{s}_t$ at time t of the discrete-time process that is governed by the linear equation

$$\mathbf{s}_t = \mathbf{A}\mathbf{s}_{t-1} + \mathbf{B}\mathbf{u}_t + \mathbf{w}_t \quad (4)$$

with a measurement $\mathbf{z}_t$, given by

$$\mathbf{z}_t = \mathbf{H}\mathbf{s}_t + \mathbf{v}_t \quad (5)$$

In (4), $\mathbf{A}$ is the state transition matrix applied to the previous state $\mathbf{s}_{t-1}$, $\mathbf{B}$ is the control input matrix that is applied to the control vector $\mathbf{u}_t$ at time t and $\mathbf{w}_t$ is a random vector that represents process noise. In (5), $\mathbf{H}$ is the observation matrix which maps the true state to the observed state and $\mathbf{v}_t$ is a random vector that represents measurement noise. $\mathbf{w}_t$ is assumed to be drawn from a zero mean multivariate normal distribution with covariance $\mathbf{Q}$, referred to as the process noise covariance. Similarly, $\mathbf{v}_t$ is modeled using a zero mean Gaussian with covariance $\mathbf{R}$, the measurement noise covariance. $\mathbf{w}_t$ and $\mathbf{v}_t$ are assumed to be independent of each other while $\mathbf{Q}$ and $\mathbf{R}$ are assumed to be constant over the entire sequence.

The Kalman filter keeps track of the state estimate $\hat{\mathbf{s}}_t$ as well as the state estimate error covariance $\hat{\mathbf{P}}_t$. During the *prediction* stage, these values ($\hat{\mathbf{s}}_t^-$ and $\hat{\mathbf{P}}_t^-$) are obtained using (6) and (7).

$$\hat{\mathbf{s}}_t^- = \mathbf{A}\hat{\mathbf{s}}_{t-1} + \mathbf{B}\mathbf{u}_{t-1} \quad (6)$$

$$\hat{\mathbf{P}}_t^- = \mathbf{A}\hat{\mathbf{P}}_{t-1}\mathbf{A}^T + \mathbf{Q} \quad (7)$$

During the *correction* stage, a *Kalman gain* $\mathbf{K}$ is computed using (8), which minimizes the *a posteriori* error covariance, and is subsequently used to refine the predicted state estimate and error covariance using the noisy measurements and hence produce corrected values ($\hat{\mathbf{s}}_t$ and $\hat{\mathbf{P}}_t$) using (9) and (10) respectively.

$$\mathbf{K}_t = \hat{\mathbf{P}}_t^- \mathbf{H}^T (\mathbf{H}\hat{\mathbf{P}}_t^- \mathbf{H}^T + \mathbf{R})^{-1} \quad (8)$$

$$\hat{\mathbf{s}}_t = \hat{\mathbf{s}}_t^- + \mathbf{K}_t(\mathbf{z}_t - \mathbf{H}\hat{\mathbf{s}}_t^-) \quad (9)$$

$$\hat{\mathbf{P}}_t = (\mathbf{I} - \mathbf{K}_t\mathbf{H})\hat{\mathbf{P}}_t^- \quad (10)$$

3 Related Work

ASMs and Kalman filters have been used in conjunction for tracking of different objects in video sequences, most commonly, human contours. Baumberg and Hogg [18] describe a method for such an application by tracking the shape coefficients of an ASM independently and attain a speed up in fitting due to this. Baumberg builds on this work and in [19] proposes a more efficient tracking method that utilizes knowledge gained from the Kalman filtering process in order to not only initialize the ASM but also improve the Mahalanobis search direction when determining the best candidate location for a landmark. Lee *et al.* [20] achieve real time tracking of human contours using a hybrid algorithm which predicts the initial human outline using Kalman filtering in combination with block matching and a hierarchical ASM to perform model fitting.

When used for tracking facial landmarks, Pu *et al.* [21] report results obtained using an ASM in combination with a Meanshift [22] method and a Kalman filter to obtain a bounding box around the face and hence initialize the ASM in every frame. Our approach differs from that of [21] because we use Kalman filters to track individual landmarks instead of using them to just track the face and initialize the ASM. This way we harness more information from the filtering process and rely less on the ASM producing very accurate fitting results as [21] does. Ahlberg [23] proposes a near real-time face tracking method that uses an Active Appearance Model [12], [24]. We use ASMs and not AAMs since the former are faster and less affected by illumination variations [12] than the latter and are hence preferred when accurate fitting of facial landmarks is the goal.

4 Tracking Methods

In this section we describe five different tracking methods that we will be comparing using experiments described in section 5.

Fig. 2. Sequence of steps involved in our proposed landmark tracking methods.

4.1 Approaches Based Only on ASM Fitting

The first three tracking methods we evaluate are based purely on ASM fitting and do not use a Kalman filtering component. An implementation that treats each frame independently is the simplest scheme that can be used. Such a method requires face detection on every frame to initialize the ASM and fails when a face is not detected or is detected in the wrong location. For the former case, we simply scale the mean shape (that was obtained from the training stage) according to the test image dimensions and place it over the test image to obtain the final landmark coordinates. Predictably, this method exhibits extremely large fitting errors whenever initialization is poor.

The next method again treats each frame independently but we manually initialize the ASM by providing it with the location of the face in frames where the face detection step failed in order to understand how much better the ASM could have performed had it been initialized well every time.

Our third method utilizes temporal information by initializing the ASM on the current frame using the fitting results of the previous frame. It is assumed that the first frame is fitted with reasonable accuracy and that the motion of the face between frames is sufficiently small. This method achieves acceptable results but is still completely reliant on high ASM fitting accuracy.

4.2 Kalman Filtered ASM Approaches

The fitting results of an ASM can not be completely trusted especially if it hasn't been initialized well. Hence, the method of using the fitting results of the previous frame to initialize the ASM for the current frame is prone to error. An improved approach is to utilize a Kalman filter to obtain optimal state estimates of the landmark positions in each frame. The prediction step of the Kalman filter

gives an *a priori* estimate of the landmark locations in the next frame and thus provides extremely good initialization. The landmark coordinates produced by the ASM as a results of this initialization are treated as noisy observations to the Kalman filter and are corrected to produce optimal, smoothed state estimates which serve as the final landmark coordinates. The process is illustrated in Fig. 2.

In our first Kalman filtering approach, we use a *constant acceleration model*, described in [25], to track the locations (x^i, y^i) and velocities (v_x^i, v_y^i) of each of the 79 facial landmarks as it has proven to be effective for visual tracking. In this model, the state vector $\mathbf{s}_t^i = [x^i \ y^i \ v_x^i \ v_y^i \ a_x^i \ a_y^i]^T$ includes the accelerations of the landmarks (a_x^i, a_y^i) as well. The state transition matrix $\mathbf{A}$ and observation matrix $\mathbf{H}$ are defined as

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & \Delta t & 0 & \frac{1}{2}\Delta t^2 & 0 \\ 0 & 1 & 0 & \Delta t & 0 & \frac{1}{2}\Delta t^2 \\ 0 & 0 & 1 & 0 & \Delta t & 0 \\ 0 & 0 & 0 & 1 & 0 & \Delta t \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{H} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix} \quad (11)$$

where Δt represents the duration of a time step.

The control input matrix $\mathbf{B}$ is ignored since there is no external contribution in our setup. The process and measurement noise covariances, $\mathbf{Q}$ and $\mathbf{R}$, are given by

$$\mathbf{Q} = \begin{bmatrix} \frac{1}{4}\Delta t^4 & 0 & \frac{1}{2}\Delta t^3 & 0 & \frac{1}{2}\Delta t^2 & 0 \\ 0 & \frac{1}{4}\Delta t^4 & 0 & \frac{1}{2}\Delta t^3 & 0 & \frac{1}{2}\Delta t^2 \\ \frac{1}{2}\Delta t^3 & 0 & \Delta t^2 & 0 & \Delta t & 0 \\ 0 & \frac{1}{2}\Delta t^3 & 0 & \Delta t^2 & 0 & \Delta t \\ \frac{1}{2}\Delta t^2 & 0 & \Delta t & 0 & 1 & 0 \\ 0 & \frac{1}{2}\Delta t^2 & 0 & \Delta t & 0 & 1 \end{bmatrix} \sigma_v^2, \quad \mathbf{R} = \begin{bmatrix} \phi & 0 \\ 0 & \phi \end{bmatrix} \quad (12)$$

where the values of σ_v^2 and ϕ were empirically estimated. σ_v^2 was set to 3 and ϕ was set to 10 for landmarks around the face boundary and 4 for all remaining points, since the landmarks along the facial boundary show relatively large fitting errors and hence have a greater deal of uncertainty associated with their positions.

A possible drawback of this scheme is that it does not account for the correlated motion of the facial landmarks. In order to take advantage of this, we designed another tracking scheme, in which rather than tracking the landmark locations themselves, we track parameters that affect these positions. These consist of the translation of the mean of all 79 points in the image (t_x and t_y), the rotation of the face (θ), the size of the face (r), and the first four PCA coefficients of the facial structure ($b_1 - b_4$). Only four PCA coefficients were tracked as the remaining coefficients did not demonstrate smooth trajectories in our tests and also model a very small percentage of the variance associated with the different facial shapes in the training set. Empirical results also showed that constraining

Table 1. Performance of the different tracking methods on three video sequences. The best performance values for each video are shown in bold and demonstrate the effectiveness of our methods that utilize Kalman filtering.

Method	Video 1 (120 Frames)		Video 2 (100 Frames)		Video 3 (70 Frames)	
	Fit Error Mean (pixels)	Fit Error Std. Devn. (pixels)	Fit Error Mean (pixels)	Fit Error Std. Devn. (pixels)	Fit Error Mean (pixels)	Fit Error Std. Devn. (pixels)
ASM on Individual Frames	17.86	31.99	10.10	16.01	9.27	15.50
ASM on Individual Frames with Correction	10.25	9.68	7.18	3.83	6.78	4.98
ASM using Previous Frame Results	8.80	6.17	10.55	16.06	6.21	1.79
ASM with Kalman Filtering of Landmark Posns.	7.58	3.59	6.43	1.97	6.19	1.76
ASM with Kalman Filtering of PCA Shape Coeffs.	7.55	3.67	6.44	2.07	6.54	2.08

a larger number of PCA coefficients using the Kalman filter resulted in higher fitting errors. Using estimates of the tracked components and the remaining PCA coefficients, we could reconstruct the coordinates of all the landmarks. *Constant velocity models* [25] were used for tracking the translation, size of the face and PCA coefficients and a *constant angular velocity model* [25] was used for modeling the rotation.

5 Experiments and Results

We evaluate the different landmark tracking methods on frames from three video sequences that were recorded in a lab setting and consist of a single person exhibiting movement of the face along all three axes. The videos have 120, 100 and 70 frames respectively, with a resolution of 640×480 pixels. Ground truth landmark locations for each of the frames in the video sequences were manually annotated.

By calculating the mean of the Euclidean distances between the coordinates of each of the 79 landmarks that were fitted by the tracking method to those that were manually annotated, we obtain the fitting error for a particular frame. The mean and standard deviation of this fitting error across all frames in a video sequence provide us with performance parameters that can be used to

Fig. 3. Examples of ASM initialization on frame 18 of video 1 using different methods (a) Initialization provided by face detection (b) Initialization provided by using fitting results of the previous frame (c) Initialization using the prediction step of a Kalman filter built to track landmark coordinates.

quantitatively compare different methods. We interpret the mean fitting error as a measure of accuracy, and the standard deviation as a measure of consistency of the approach.

The results obtained by the different landmark tracking implementations on the three videos using these performance parameters are shown in Table 1. It is clear that the performance of the naïve ASM implementation that fits each frame independently is the worst across all videos. This is expected since the frames where the face detection was poor caused large errors and also caused the standard deviation of the fitting error to grow. We found that even though the data was collected under carefully controlled lab conditions, with a plain white background, face detection still failed on approximately 10% of the frames. By manually providing the correct location of the face to the ASM in the frames where initialization was extremely poor we obtain better results.

When the ASM is initialized using the fitting results of the previous frame, the fitting accuracy is further increased and the standard deviation of the fitting error also falls. This method performed well on videos 1 and 3 but not on video 2 as it lost track of the face between frames 62 and 68 and as a result of this ended up with large values of mean and standard deviation of fitting error.

Our implementation that tracks the positions of landmarks using a Kalman filter performs extremely well and obtains a mean fitting error across all videos that is lower than all the previous methods. Clearly, the predictive-corrective mechanism used by the Kalman filtering process enables better initialization and also refines the final landmark positions once the ASM fitting process has been completed. Results obtained on videos 1 and 2 best illustrate this as our method is clearly far more accurate than the schemes that do not utilize Kalman fitting. On video 3 our approach produces results that are quite close to the results obtained by simply initializing the ASM using the previous frame. This was because the pose and in-plane rotation of the face in this video were slightly less than in the first two videos allowing the ASM fitting results to be trusted more and hence limiting the role played by Kalman filtering.

Fig. 4. Landmark fitting results obtained by running different methods on frames from our video sequences. Each column of images corresponds to fitting results on frames 18, 76 and 89 from video 1, frames 12, 64 and 68 from video 2 and frames 17, 36 and 43 from video 3 respectively (a) Method 1 - ASM run on individual frames (b) Method 2 - ASM run on individual frames with correct initialization provided for frames that suffered from poor face detection (c) Method 3 - ASM run on each frame using landmark coordinates in the previous frame for initialization (d) Method 4 - Kalman filtering of landmark coordinates used in combination with ASM (e) Method 5 - Kalman filtering of PCA coefficients of face shapes used in combination with ASM.

Fig. 5. Graphs showing the mean fitting error obtained by the different tracking methods on landmarks 1-79 for frames from three videos (a) video 1 (b) video 2 (c) video 3.

Our last tracking method that uses Kalman tracking of the PCA facial shape coefficients instead of tracking the individual landmark coordinates, also obtained extremely accurate results on all three videos. There was not much to choose from between both the Kalman based approaches as the landmark coordinate tracking approach performed slightly better on videos 2 and 3, while the method of tracking facial shape coefficients fared marginally better on video 1.

Fig. 3 compares the initialization provided to the ASM on a particular frame as a result of using three different methods and shows that the Kalman based approaches obtain far better initialization which plays a major role in improving their landmark fitting accuracy. Fig. 4 shows samples of the fitting results obtained by the different methods on three frames from each video. It is clear that both our Kalman based methods obtain accurate fitting results especially on the eyes, eyebrows and lips. The nose and facial outline are difficult to fit since the ASM has been trained only on frontal images and does not handle large pose variation well, however, our tracking methods do an acceptable job

with these landmarks too when compared to the purely ASM based methods. Fig. 4 also shows how the most basic method that simply fits each frame separately is thrown off completely when face detection fails to locate the face or provides incorrect initialization. Even when correct initialization is provided to this method, it does not accurately fit faces exhibiting large in-plane rotation or pose variation. The ASM method that utilizes fitting results of the previous frame does better than the previous two methods but fails on frames 62 - 68 of video 2 and this serves in best illustrating why Kalman filtering is so important.

Fig. 5 shows how the fitting error of each method varies for the different landmarks when averaged over all the frames in videos 1, 2 and 3 respectively. The Kalman based tracking methods obtain lower fitting error for all landmarks than the other methods but even they show high fitting errors when dealing with points along the facial outline (landmarks 1-15) and along the eyebrows (landmarks 52-65). The best fitted points for all the methods lie along the eyes (landmarks 16-31) as they are regions with easily distinguishable local pixel intensity gradients and also because the inherent variability in human landmarking is lowest for these points. On video 3, except for the most basic approach which treats each frame independently, the remaining methods produce very similar results. Hence, the fitting error curves overlap for these methods and thus there is not much difference between the mean and standard deviation of the fitting errors across all frames produced by these approaches, which can be seen in Table 1. It is worth noting that for each video the shapes of all fitting error curves obtained by the five methods are quite similar and are just offset with respect to each other, indicating the inherent difficulty in fitting certain landmarks over others.

6 Conclusions and Future Work

We have provided a thorough analysis of several methods that could be used to track facial landmarks across frames of a video sequence by testing them on frames from three videos. However accurate an ASM implementation may be, an approach that treats each frame independently and does not utilize temporal information is fraught with risk as it will perform very poorly on every frame where the face detection results are off. We have proposed two tracking implementations that utilize Kalman filters to refine the ASM fitting results. Our first method tracks the individual landmark positions using a Kalman filter for each one while our second method tracks several parameters including the PCA shape coefficients of the landmark coordinates and uses these parameters to obtain the positions of the landmarks. Both methods outperform a scheme that just utilizes the results of the previous frame to initialize an ASM on the current frame as the predictive mechanism of the Kalman filtering process helps in initializing the ASM better while the corrective mechanism refines its search results. The combination of the two mechanisms ensures that both our approaches are less reliant on perfect fitting results from the ASM and are able to consistently provide accurate results.

The methods we have proposed currently work on the assumption that there is a single face in the video that needs to be tracked and that there is no abrupt scene change, change in the camera position or zooming in of the subject. Future work to deal with videos that exhibit any of these changes between frames would require background subtraction and re-initialization of the ASM using face detection. The use of several ASMs each trained on a different pose would help in boosting fitting accuracy as would the use of a rotation-tolerant face detection system. Our ASM implementation takes a few seconds to fit an image and hence our methods have currently been designed for post-processing of videos as they are not yet fast enough for real-time applications. Thus, another area of future work would involve speeding up of our ASM implementation as well as the Kalman tracking process to build a system capable of operating on surveillance videos in real-time. Finally, we also need to benchmark our approaches against contemporary landmark tracking methods on publicly available datasets as well as on challenging video sequences containing more natural (i.e. cluttered) backgrounds with subjects exhibiting faster motion of the head and lip movements due to varied expressions and speech.

Acknowledgments. We would like to thank the U.S. Army Research Lab and Carnegie Mellon CyLab for partially funding this work.

References

1. Cootes, T.F., Hill, A., Taylor, C.J., Haslam, J.: The Use of Active Shape Models for Locating Structures in Medical Images. *Image and Vision Computing*. 12, 355–366 (1994)
2. de Bruijne, M., van Ginneken, B., Viergever, M.A., Niessen, W.J.: Adapting Active Shape Models for 3D Segmentation of Tubular Structures in Medical Images. In: *Information Processing in Medical Imaging*, pp. 136–146. Springer-Verlag, Berlin, Heidelberg (2003)
3. Cootes, T.F., Taylor, C.J., Lanitis, A.: Active Shape Models : Evaluation of a Multi-Resolution Method for Improving Image Search. In: *Proceedings of the 5th British Machine Vision Conference*, pp. 327–336 (1994)
4. Milborrow, S., Nicolls, F.: Locating Facial Features with an Extended Active Shape Model. In: *ECCV '08: Proceedings of the 10th European Conference on Computer Vision: Part IV*, pp. 504–513. Springer-Verlag, Berlin, Heidelberg (2008)
5. Seshadri, K., Savvides, M.: Robust Modified Active Shape Model for Automatic Facial Landmark Annotation of Frontal Faces. In: *BTAS '09: Proceedings of the 3rd IEEE International Conference on Biometrics: Theory, Applications and Systems*, pp. 319–326. IEEE Press, Piscataway, NJ, USA (2009)
6. Gu, L., Kanade, T.: A Generative Shape Regularization Model for Robust Face Alignment. In: *ECCV '08: Proceedings of the 10th European Conference on Computer Vision: Part I*, pp. 413–426. Springer-Verlag, Berlin, Heidelberg (2008)
7. Edwards, G.J., Cootes, T.F., Taylor, C.J.: Face Recognition Using Active Appearance Models. In: *ECCV '98: Proceedings of the 5th European Conference on Computer Vision, Volume 2*, pp. 581–595. Springer-Verlag, Berlin, Heidelberg (1998)

8. Choi, H., Oh, S.: Real-time Recognition of Facial Expression Using Active Appearance Model with Second Order Minimization and Neural Network. In: Proceedings of the 2006 IEEE International Conference on Systems, Man, and Cybernetics, pp. 1559–1564 (2006)
9. Zhao, S., Gao, Y.: Automated Face Pose Estimation Using Elastic Energy Models. In: ICPR '06: Proceedings of the 18th International Conference on Pattern Recognition - Volume 04, pp. 618–621. IEEE Computer Society, Washington, DC, USA (2006)
10. Kalman, R.E.: A New Approach to Linear Filtering and Prediction Problems. Transactions of the ASME – Journal of Basic Engineering. 82, 35–45 (1960)
11. Gower, J.C.: Generalized Procrustes Analysis. Psychometrika. 40, 33–51 (1975)
12. Cootes, T.F., Taylor, C.J.: Statistical Models of Appearance for Computer Vision. Technical report, Imaging Science and Biomedical Engineering, University of Manchester (2004)
13. Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J.: Active Shape Models - Their Training and Application. Computer Vision and Image Understanding. 61, 38–59 (1995)
14. Intel: Open Source Computer Vision Library, <http://software.intel.com/sites/oss/pdfs/OpenCVreferencemanual.pdf> (2007)
15. Viola, P., Jones, M.J.: Robust Real-Time Face Detection. International Journal of Computer Vision. 57, 137–154 (2004)
16. Phillips, P.J., Flynn, P.J., Beveridge, J.R., Scrugs, W.T., O Toole, A.J., Bolme, D., Bowyer, K.W., Draper, B.A., Givens, G.H., Lui, Y.M., Sahibzada, H., Scallan III, J.A., Weimer, S.: Overview of the Multiple Biometrics Grand Challenge. In: Proceedings of the 3rd IAPR/IEEE International Conference on Biometrics, pp. 705–714. Springer-Verlag, Berlin, Heidelberg (2009)
17. National Institute of Standards and Technology (NIST): Multiple Biometric Grand Challenge (MBGC) - 2008, <http://face.nist.gov/mbgc> (2008)
18. Baumberg, A.M., Hogg, D.C.: An Efficient Method for Contour Tracking using Active Shape Models. In: Proceedings of the 1994 IEEE Workshop on Motion of Nonrigid and Articulated Objects, pp. 194–199 (1994)
19. Baumberg, A.: Hierarchical shape fitting using an iterated linear filter. Image and Vision Computing. 16, 329–335 (1998)
20. Lee, S.W., Kang, J., Shin, Paik, J.: Hierarchical active shape model with motion prediction for real-time tracking of non-rigid objects. IET Computer Vision. 1, 17–24 (2007)
21. Pu, B., Liang, S., Xie, Y., Yi, Z., Heng, P.: Video Facial Feature Tracking with Enhanced ASM and Predicted Meanshift. In: Proceedings of the 2nd International Conference on Computer Modeling and Simulation, Volume 2, pp. 151–155. IEEE Computer Society, Los Alamitos, CA, USA (2010)
22. Comaniciu, D., Ramesh, V., Meer, P.: Real-Time Tracking of Non-Rigid Objects using Mean Shift. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Volume 2, pp. 142–149. IEEE Computer Society, Los Alamitos, CA, USA (2000)
23. Ahlberg, J.: An Active Model for Facial Feature Tracking. EURASIP Journal on Applied Signal Processing. 2002, 566–571 (2002)
24. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active Appearance Models. In: ECCV '98: Proceedings of the 5th European Conference on Computer Vision, Volume 2, pp. 484–498. Springer-Verlag, Berlin, Heidelberg (1998)
25. Bar-Shalom, Y., Kirubarajan, T., Li, X.R.: Estimation with Applications to Tracking and Navigation. John Wiley & Sons, Inc., New York, NY, USA (2002)