Facial Landmark Tracking In Videos Using Kalman Filter Assisted Active Shape Models

Utsav Prabhu (uprabhu@andrew.cmu.edu)
Keshav Seshadri (kseshadr@andrew.cmu.edu)
Prof. Marios Savvides (msavvid@ri.cmu.edu)
Carnegie Mellon University
Overview

- Motivation
- Background
  - Active Shape Model
  - Discrete Kalman Filter
- Methods Compared
  - Purely ASM Based Methods
  - Kalman Filter Assisted ASMs
- Results
- Conclusions
- Future Work
Motivation

- Locating facial landmarks across video frames can aid
  - Facial Recognition
  - Pose Correction
  - Expression Analysis

- An Active Shape Model (ASM) can be used for this purpose

- Use of an ASM on individual frames is prone to error

- Kalman filters allow for refinement of ASM results and better initialization on next frame
Active Shape Models (ASMs) [1]

Generate facial model using training images

Detect face in test image

Deform model to fit face in test image

ASM Training Stage – Facial Shape

- 79 facial landmarks manually marked on training set (4000 images from still challenge set of MBGC – 2008[2])

- Our landmarking scheme

- Training set samples

- Apply PCA to shapes (aligned using Procrustes analysis[3])

- Shape model equation governs coordinates of landmarks in any new shape

\[ \mathbf{x} \approx \overline{\mathbf{x}} + \mathbf{Pb} \]

- \( \mathbf{X} \) - Any facial shape
- \( \overline{\mathbf{x}} \) - Mean shape
- \( \mathbf{P} \) - Eigenvectors matrix
- \( \mathbf{b} \) - Projection coefficients


ASM Training Stage – Profiling

- 2D profiles (of image gradients) built for each landmark at L (= 4) pyramid levels
- Mean profile vector ($\bar{g}$) and covariance matrix ($S_g$) calculated for each landmark

ASM Testing Stage

Face detected, mean shape aligned with face

Each landmark point moved to candidate that gives lowest Mahalanobis distance

Mahalanobis distance $f(g)$ between candidate profile and mean profile

\[ f(g) = (g - \bar{g})^T S_g^{-1} (g - \bar{g}) \]

- $g$ - Profile around candidate point
- $\bar{g}$ - Mean profile for landmark
- $S_g$ - Covariance matrix for landmark

Generate shape coefficients vector $(b)$

Constrain $b_i$ so that $|b_i| \leq 3\sqrt{\lambda_i}$

$\lambda_i$ - $i^{th}$ eigenvalue corresponding to $b_i$

Final landmark coordinates at highest resolution image

Multi-Resolution Search
Discrete Kalman Filter \([5]\)

- Predictive-corrective algorithm

- Estimate optimal state \(s_t\) at time \(t\)
  \[s_t = A s_{t-1} + B u_{t-1} + w_{t-1}\]
  with a measurement given by
  \[z_t = H s_t + v_t\]

- Prediction Stage
  - \(\hat{s}_{t}^- = A \hat{s}_{t-1} + B u_{t-1}\)
  - \(\hat{P}_{t}^- = A \hat{P}_{t-1} A^T + Q\)

- Correction Stage
  - \(K_t = \hat{P}_{t}^- H^T (H \hat{P}_{t}^- H^T + R)^{-1}\)
  - \(\hat{s}_t = \hat{s}_{t}^- + K_t (z_t - H \hat{s}_{t}^-)\)
  - \(\hat{P}_t = (I - K_t H) \hat{P}_{t}^-\)

- \(A\) – State transition matrix
- \(B\) – Control input matrix
- \(u_t\) – Control vector
- \(H\) – Observation matrix
- \(w_t\) – Process noise \(N(0, Q)\)
- \(v_t\) – Measurement noise \(N(0, R)\)
- \(\hat{s}_t\) – State estimate
- \(\hat{P}_t\) – State estimate error covariance
- \(K_t\) – Kalman Gain

Purely ASM Based Approaches

- **ASM on individual frames**
  - Does not harness temporal information
  - Results in large MSE when face detection results are off

- **ASM on individual frames with correction**
  - Manual initialization for frames where face detection results were incorrect
  - Shows best case performance of ASM on individual frames

- **ASM with initialization using previous frame**
  - Initializes ASM on next frame using previous frame results
  - Acceptable results, but highly dependent on ASM fitting
  - Can’t correct for poor fitting results on a frame, which can affect fitting on future frames
Kalman Filter Assisted ASM

**Kalman filter prediction**

\[
\hat{s}_t^- = A\hat{s}_{t-1}^- \\
\hat{P}_t^- = A\hat{P}_{t-1}^-A^T + Q
\]

**Kalman filter correction**

\[
K_t = \hat{P}_t^-H^T(H\hat{P}_t^-H^T + R)^{-1} \\
\hat{s}_t = \hat{s}_t^- + K_t(z_t - H\hat{s}_t^-) \\
\hat{P}_t = (I - K_tH)\hat{P}_t^-
\]

Face detection + ASM fitting on first frame

Video frames

Final landmark locations

\[x = H\hat{s}_t\]

Predicted landmark coordinates

\[x = H\hat{s}_t^-\]

ASM fitted landmarks (observations)

ASM initialization

ASM fitting process
Kalman Filter Assisted ASM

- Tracking landmark coordinates across frames
  - Constant acceleration model \cite{6} to track coordinates and velocities of 79 landmarks
  - Measurement noise covariance \( R \) used to account for larger variance in motion of facial boundary points

- Tracking parameters that affect landmark positions
  - Accounts for correlated motion of landmarks
  - Track translation of landmarks, rotation of face, size of face and the first four PCA facial shape coefficients
  - Constant velocity models \cite{6} for tracking translation, size of face and PCA coefficients and constant angular velocity model \cite{6} for rotation

Improvement in Initialization

Comparing initialization provided by different methods on frame 18 of video 1

- a - Initialization provided by face detection
- b - Initialization provided by using ASM results of previous frame
- c - Initialization provided by prediction step of Kalman filter
Results – Fitting Video Frames

- Video 1: Frame 89
  a – ASM on individual frames
  b – ASM on individual frames with correction
  c – ASM initialized using results of previous frame
  d – ASM with Kalman filtering of landmark coordinates
  e – ASM with Kalman filtering of parameters affecting landmark locations

- Video 2: Frame 12

- Video 3: Frame 43
# Results – Fitting Accuracy

<table>
<thead>
<tr>
<th>Method</th>
<th>Video 1 (120 Frames)</th>
<th>Video 2 (100 Frames)</th>
<th>Video 3 (70 Frames)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Fit Error Mean (pixels)</td>
<td>Fit Error Std. Devn. (pixels)</td>
<td>Fit Error Mean (pixels)</td>
</tr>
<tr>
<td>ASM on Individual Frames</td>
<td>17.86</td>
<td>31.99</td>
<td>10.10</td>
</tr>
<tr>
<td>ASM on Individual Frames with Correction</td>
<td>10.25</td>
<td>9.68</td>
<td>7.18</td>
</tr>
<tr>
<td>ASM using Previous Frame Results</td>
<td>8.80</td>
<td>6.17</td>
<td>10.55</td>
</tr>
<tr>
<td>ASM with Kalman Filtering of Landmark Positions</td>
<td>7.58</td>
<td>3.59</td>
<td>6.43</td>
</tr>
<tr>
<td>ASM with Kalman Filtering of PCA Shape Coefficients</td>
<td>7.55</td>
<td>3.67</td>
<td>6.44</td>
</tr>
</tbody>
</table>

Performance of 5 different tracking methods on three video sequences
Conclusions

- Comparison of several methods to track facial landmarks in videos
- Purely ASM based methods are naïve and seldom work well
- Proposed two Kalman filter based schemes
  - Kalman filters for tracking individual landmark coordinates
  - Kalman filters for tracking parameters that affect landmark positions
- Experiments on 3 videos confirm our Kalman based approaches enable better ASM initialization and lower fitting errors
Future Work

- Background subtraction and re-initialization of ASM to deal with scene changes, zooming in of subject etc

- Speed optimizations for our ASM and Kalman tracking implementations

- Benchmark our approach on publicly available datasets/more challenging datasets
Acknowledgments

We would like to thank:

Carnegie Mellon CyLab
U.S. Army Research Lab
Questions ?