Discriminative Pose Tracking with Mixtures of Gaussian Processes

Martin Fergie and Aphrodite Galata

School of Computer Science, University of Manchester

3rd September 2012
Discriminative Pose Estimation

\[ \mathbf{X} \xrightarrow{f(x)} \mathbf{Y} \]
Challenges

- Large data sets,
- High dimensionality,
- Multi-modal,
- Large amounts of ambiguity.
Contributions

- Mixture of GP Appearance Model
  
- 2nd Order Dynamics Framework
Mixture of Experts Model

Weighted combination of probabilistic regression functions

\[ y = \sum_i \pi_i f_i(x) \] (1)

\[ p(y|x, \theta) = \sum_i p(i|x, \theta_g)p(y|x, \theta_i) \] (2)

Each \( f(x) \) models a local region of the data set

\[ p(y|x, \theta) = \sum_i \underbrace{p(i|x, \theta_g)\mathcal{N}(y|\mu_i(x), \Sigma_i(x))}_{\text{posterior}} \] (3)

\[ \underbrace{\mathcal{N}(y|\mu_i(x), \Sigma_i(x))}_{\text{expert prediction}} \]
Gaussian Processes

Gives a joint Gaussian distribution over a set of random variables

\[ p(y_*|x_*, X, Y, \theta) = \mathcal{N}(y_*|\mu_i(x_*), \Sigma_i(x_*) ) \]  \hspace{1cm} (4)
Applying Gaussian Processes into a Mixture of Experts Model

Introduce a hard partitioning of the data set using indicator variable $z$

$$p(y_*|x_*, z) = \sum_{i=1}^{K} p(z_n = i|x_*, \theta_g) p(y|x_*, \mathbf{X}_{\vartheta_i}, \mathbf{Y}_{\vartheta_i}, \theta_i).$$  \hspace{1cm} (5)

where

$$\vartheta_i = \{n|n \in N, z_n = i\}$$  \hspace{1cm} (6)
Optimising the Expert Indicators

Optimise $z$ using Gibbs sampling

$$p(z_n = i | z_{/n}, X, Y, \theta_i, \theta_g) \propto p(y_n | x_n, X_{\theta_i/n}, Y_{\theta_i/n}, \theta_i) p(z_n = i | x_n, \theta_g).$$

Video: Gibbs sampling process
Problem: How to infer a smooth sequence of pose estimates \( \hat{Y} = \{\hat{y}_n\} \) from a series of Gaussian mixture distributions.

Naive Approach: \( \hat{y}_n = \mathbb{E}[p(y_n|x_n)] \)

- Averages out multi-modality,
- No temporal dependency on previous frames, jittery tracking.
Incorporating a Dynamics Constraint

\[
p(y_n | x_n, y_{1:n-1}) = p(y_n | x_n) \cdot p(y_n | y_{1:n-1})
\]

\[\text{Combined Prediction} \quad \text{Appearance Model} \quad \text{Dynamics Model} \quad (7)\]

Challenges

- Dynamics model is sensitive to previous pose estimates.
- Mixture of Gaussian predictive distribution is not compatible with standard LDS type models.
Dynamical Pose Filtering

- Maintain a prediction for each expert at each frame,
- Use $z_n$ to break the mixture of Gaussians into independent predictions,
- Combine with a Gaussian second order dynamical model,
- Infer optimal sequence of $z = \{z_n\}$.
Second Order Dynamics

$$y_{n-1} \rightarrow y_n$$

$$p(y_n|y_{n-1}, y_{n-2}) = \mathcal{N}(\mu([y_{i,n-1}, y_{i,n-2}]^T), \sigma([y_{i,n-1}, y_{i,n-2}]^T))$$

(8)
Inferring the Optimal Observation Sequence

Prediction of expert $i$ at frame $n$

$$\hat{y}_{i,n} = p(y_n, z_n = i | x_{1:n}) = p(y_n | x_n, z_n) \sum_{z_{n-2}} \sum_{z_{n-1}} p(y_n | \hat{y}_{z_{n-1,n-1}, n-1}, \hat{y}_{z_{n-2,n-2}}) p(z_{n-1} | x_{1:n-1}) p(z_{n-2} | x_{1:n-2})$$

(9)
Inferring the Optimal Observation Sequence

Prediction of expert $i$ at frame $n$

\[
\hat{y}_{i,n} = p(y_n, z_n = i | x_{1:n}) = \left\{ p(y_n | x_n, z_n) \right\} \sum_{z_{n-2}} \sum_{z_{n-1}} p(y_n | \hat{y}_{z_{n-1},n-1}, \hat{y}_{z_{n-2},n-2}) p(z_{n-1} | x_{1:n-1}) p(z_{n-2} | x_{1:n-2})
\]  

(10)
Inferring the Optimal Observation Sequence

Prediction of expert $i$ at frame $n$

$$\hat{y}_{i,n} = p(y_n, z_n = i | x_{1:n}) = p(y_n | x_n, z_n)$$

$$\sum_{z_{n-1}} \sum_{z_{n-2}} p(y_n | \hat{y}_{z_{n-1}, n-1}, \hat{y}_{z_{n-2}, n-2}) p(z_{n-1} | x_{1:n-1}) p(z_{n-2} | x_{1:n-2})$$

(11)
Inferring the Optimal Observation Sequence

Node marginals

\[
p(z_n|x_{1:n}) = \underbrace{p(z_n|x_n)}_{\text{appearance gating}} \underbrace{p(\hat{y}_{z_n,n}|Y_{tr})}_{\text{pose density}} \\
\sum_{z_{n-2}} \sum_{z_{n-1}} \underbrace{p(z_{n-1}|x_{1:n-1})p(z_{n-2}|x_{1:n-2})}_{\text{previous marginals}},
\]

(12) \hspace{1cm} (13)
Inferring the Optimal Observation Sequence

Max-sum algorithm for Bayesian networks:

- Forward pass over sequence to evaluate predictions $y_{n,i}$ and marginals $p(z_n = i|x_{1:n})$,
- Backward pass to find optimal $z$ over full sequence.
- Pose prediction for each frame is then $\hat{y}_n = y_{n,i=z_n}$
Results on Human Pose Tracking

Results videos
Quantitative Results: Appearance Model

Performance compared to other discriminative models

<table>
<thead>
<tr>
<th>Dataset: Model/Feature</th>
<th>Ballet</th>
<th>Sign Language</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BOW SC</td>
<td>HMAX</td>
</tr>
<tr>
<td>Our Method (app)</td>
<td>32.52</td>
<td>32.68</td>
</tr>
<tr>
<td>BME ¹</td>
<td>51.71</td>
<td>71.72</td>
</tr>
<tr>
<td>Urtasun and Darrell ²</td>
<td>36.13</td>
<td>38.18</td>
</tr>
<tr>
<td>sKIE ³</td>
<td>31.55</td>
<td>37.57</td>
</tr>
<tr>
<td>Kernel Regression</td>
<td>71.68</td>
<td>71.71</td>
</tr>
</tbody>
</table>

¹Bo and Sminchisescu CVPR 2008
²Urtasun and Darrell CVPR 2008
³Memisevic et al. PAMI 2012
Quantitative Results: Dynamics Model

<table>
<thead>
<tr>
<th>Dataset</th>
<th>DPF</th>
<th>LDS</th>
<th>Appearance Only</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sign Language</td>
<td>7.06</td>
<td>7.28</td>
<td>6.88</td>
</tr>
<tr>
<td>Ballet</td>
<td>32.19</td>
<td>50.14</td>
<td>32.37</td>
</tr>
</tbody>
</table>
Questions