In this paper we present a method for performing discriminative human pose estimation using a mixture of Gaussian Processes appearance model to map directly from the image features to the multi-model pose distribution. In order to obtain a pose estimate for a sequence of frames, we introduce a dynamic programming algorithm for inferring a smooth pose sequence from the multi-model distribution given by our appearance model.

## 1 Mixture of Gaussian Processes

A mixture of experts model gives a predictive distribution over the pose $y$ conditioned on the image observation $x$ as a mixture of Gaussian distributions:

$$p(y|x) = \sum_{i=1}^{K} p(z = i|x)\mathcal{N}(\mu_{i}(x), \Sigma_{i}(x)),$$

where each $\mu_{i}$ and $\Sigma_{i}$ are given as a function of $x$ and $p(z = i|x)$ is a weight applied to each component as a function of $x$ such that $\sum_{i} p(z = i|x) = 1$ and $0 \leq p(z = i|x) \leq 1$. We use a model where each expert prediction, $\mathcal{N}(\mu_{i}(x), \Sigma_{i}(x))$, is given by a Gaussian Process allowing each model to map a non-linear region of the dataset. To learn the model we partition the data set using an indicator variable $z = \{z_{n}\}_{n=1}^{N}$ where each $z_{n} = i$ indicates that training point $n$ is used to train expert $i$. We initialise $z$ using k-means and then use a Gibbs sampling algorithm to optimise $z$ with respect to the pose distribution:

$$p(z_{n} = i|x_{n}, X, Y, \theta_{i}, \phi) \propto p(y_{n}|x_{n}, X_{0:n}, Y_{0:n}, \theta_{i}) \cdot p(z_{n} = i|x_{n}, \phi).$$

where $z = \{z_{n}\}_{n=1}^{N}$, $z_{n} \in \{1 \ldots K\}$ indicates which expert each data point belongs to, $\theta_{i}/n$ is the set of indices of data points that belong to expert $i$, with the $n^{th}$ point removed. Learning is performed in an expectation-maximisation fashion where we iterate between training the experts and resampling $z$.

Figure 2 demonstrates the effect of the Gibbs sampling process. The top figure gives the predictive distribution when $z$ is set by running k-means on the training pose data. The lower plot shows that the Gibbs iterations leads to a more accurate pose distribution.

## 2 Dynamic Pose Filtering

![Figure 1: Graphical model for 2nd order pose filtering showing the nodes involved in computing $y_{n}$.](image)

The mixture of experts model gives us a Gaussian mixture model over the pose for each frame. The naive approach to estimating the pose for each frame would be to take the expectation of this distribution, getting a pose estimate by taking a weighted average of the Gaussian components. This averages out the multi-modal regions and does not utilise any temporal information resulting in a jittery tracking sequence.

We introduce a dynamic programming algorithm that incorporates a second order dynamical prediction model to infer a smooth path through the predictive distributions of each frame. Our algorithm propagates multiple predictions for each frame, where each prediction represents the observation of one appearance expert. For frame $n$ and expert $z_{n} = i$, we obtain a Gaussian prediction over the pose given by:

$$\hat{y}_{n} = \frac{\sum p(y_{n} | y_{n-1}, z_{n}=i) p(y_{n-1} | z_{n-1}, x_{n-1}) p(z_{n-1} | x_{n-1}) p(z_{n-2} | x_{n-2})}{\sum_{z_{n-1}} \sum_{z_{n-2}} p(y_{n} | y_{n-1}, z_{n-1}, x_{n}) p(y_{n-1} | z_{n-1}, x_{n-1}) p(z_{n-1} | x_{n-1}) p(z_{n-2} | x_{n-2})}.$$

Where $p(y_{n} | y_{n-1}, z_{n})$ is the appearance prediction for expert $z_{n} = i$ and $p(y_{n} | \hat{y}_{n-1}, z_{n}, x_{n})$ is a dynamical prediction. Figure 1 shows a graphical model illustrating this process. The forward-backward algorithm is used to infer a optimal sequence of apperance experts $z = \{z_{n}\}_{n}$ from which we obtain a smooth pose estimate $\hat{Y} = \{\hat{y}_{n}\}_{n}$.

## 3 Results

![Figure 3: Tracking results for the sign language and ballet datasets showing every fifth frame of a continuous sequence. Ground truth shown in red, predicted pose is shown in green.](image)

We evaluate our method on a 2D sign language and a 3D Ballet dance sequence. We show visual tracking results along with quantitative results comparing our method to other state of the art methods for discriminative pose estimation.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Ballet</th>
<th>Sign Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model/Feature</td>
<td>HON XX</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 [3]</td>
<td>36.13</td>
<td>38.18</td>
</tr>
<tr>
<td>sKIE [2]</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>

Table 1: Quantitative Results.