Dynamical Pose Filtering for Mixtures of Gaussian Processes

Martin Fergie
mfergie@cs.man.ac.uk
Aphrodite Galata
a.galata@cs.man.ac.uk
School of Computer Science,
University of Manchester, UK

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 $\mathbf{y}$ conditioned on the image observation $\mathbf{x}$ as a mixture of Gaussian distributions:

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

where each μ_i and Σ_i are given as a function of $\mathbf{x}$ and $p(z=i|\mathbf{x})$ is a weight applied to each component as a function of $\mathbf{x}$ such that $\sum_i p(z=i|\mathbf{x}) = 1$ and $0 \leq p(z=i|\mathbf{x}) \leq 1$. We use a model where each expert prediction, $\mathcal{N}(\mu_i(\mathbf{x}),\Sigma_i(\mathbf{x}))$, 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 $\mathbf{z} = \{z_n\}_1^N$ where each $z_n = i$ indicates that training point n is used to train expert i. We initialise $\mathbf{z}$ using k-means and then use a Gibbs sampling algorithm to optimise $\mathbf{z}$ with respect to the pose distribution:

$$p(z_n = i|\mathbf{z}_{/n},\mathbf{X},\mathbf{Y},\theta_i,\phi) \propto p(\mathbf{y}_n|\mathbf{x}_n,\mathbf{X}_{\vartheta_{i/n}},\mathbf{Y}_{\vartheta_{i/n}},\theta_i)$$
$$p(z_n = i|\mathbf{z}_{/n},\mathbf{x}_n,\phi).$$

where $\mathbf{z} = \{z_n\}_{n=1}^N$, $z_n \in \{1 \ldots K\}$ indicates which expert each data point belongs to, $\vartheta_{i/n}$ 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 $\mathbf{z}$.

Figure 2 demonstrates the effect of the Gibbs sampling process. The top figure gives the predictive distribution when $\mathbf{z}$ are 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

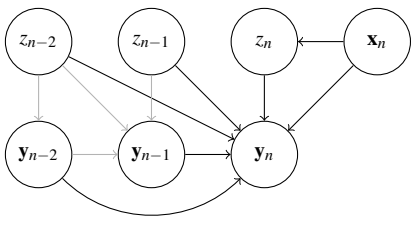

Figure 1: Graphical model for 2nd order pose filtering showing the nodes involved in computing $\mathbf{y}_n$.

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:

$$\hat{\mathbf{y}}_{i,n} = p(\mathbf{y}_n, z_n = i|\mathbf{x}_{1:n}) = p(\mathbf{y}_n|\mathbf{x}_n, z_n)$$
$$\sum_{z_{n-2}}\sum_{z_{n-1}} p(\mathbf{y}_n|\hat{\mathbf{y}}_{z_{n-1},n-1},\hat{\mathbf{y}}_{z_{n-2},n-2})p(z_{n-1}|\mathbf{x}_{1:n-1})p(z_{n-2}|\mathbf{x}_{1:n-2}).$$

Where $p(\mathbf{y}_n|\mathbf{x}_n, z_n)$ is the appearance prediction for expert $z_n = i$ and $p(\mathbf{y}_n|\hat{\mathbf{y}}_{z_{n-1},n-1},\hat{\mathbf{y}}_{z_{n-2},n-2})$ 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 $\mathbf{z} = \{z_n\}_n^N$ from which we obtain a smooth pose estimate $\hat{\mathbf{Y}} = \{\hat{\mathbf{y}}_n\}_n^N$.

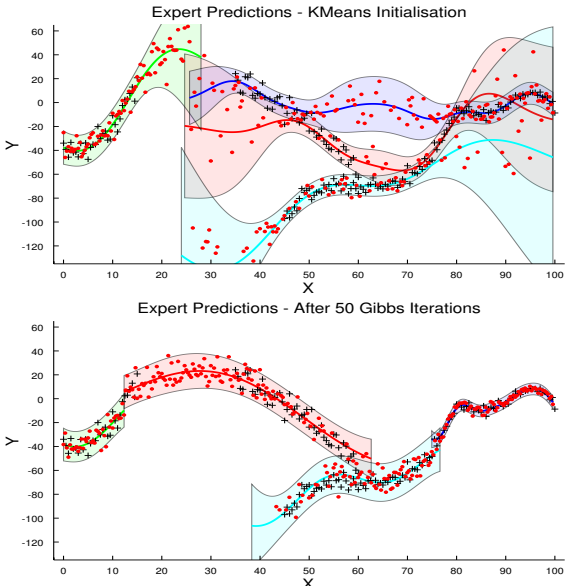

Figure 2: Predictive distributions for the mixture of Gaussian Processes model on a toy dataset. The black crosses represent the training points, the red dots are samples drawn from the predictive distribution and the coloured lines represent the predictive mean and variance of each expert.

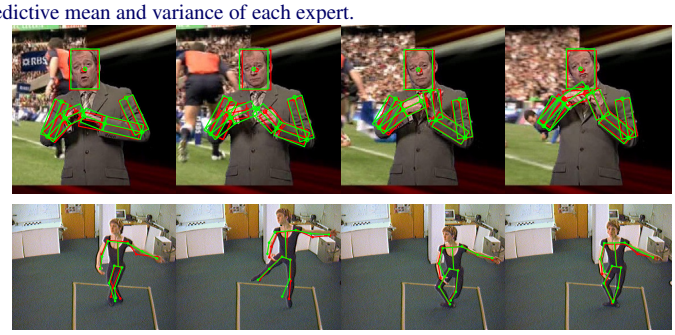

Figure 3: Tracking results for the sign language and ballet datasets showing every fifth frame of a continous sequence. Ground truth shown in red, predicted pose is shown in green.

3 Results

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

Dataset:	Ballet		Sign Language
Model/Feature	BOW SC	HMAX	HMAX
Our Method (app)	32.52	32.68	6.88 ± 0.48
BME [1]	51.71	71.72	11.87 ± 0.92
Urtasun and Darrell [3]	36.13	38.18	8.11 ± 0.62
sKIE [2]	31.55	37.57	9.36 ± 0.52
Kernel Regression	71.68	71.71	10.65 ± 0.30

Table 1: Quantitive Results.

[1] Bo and Sminchisescu. *CVPR*, 2008.
[2] Memisevic et al. *PAMI*, 2012.
[3] Urtasun and Darrell. *CVPR*, 2008.