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

Abstract

In this paper we propose a novel method for discriminative monocular human pose tracking using a mixture of Gaussian processes and a dynamic programming algorithm for selecting the optimal expert at each frame. The proposed tracking mechanism incorporates a dynamical model into the predictive distribution which is combined with the appearance model in a principled manner. This model is able to give a smoother predicted pose and resolves ambiguities in the image to pose mapping. We introduce a mixture of Gaussian processes model which optimises the size and location of each expert ensuring that each expert models a coherent region of the dataset resulting in an accurate predictive density. We compare our method to other state of the art methods on 2D and 3D monocular pose estimation on ballet and sign language data sets.

1 Introduction

Discriminative pose estimation is the task of estimating the pose of an articulated body by directly learning a mapping from image features to the subject's pose [1]. These methods use a training set containing images with the subject's pose annotated to build a model that is capable of inferring 3D pose from a single viewpoint [23].

However, the mapping from an image to its 3D pose is highly noisy and ambiguous, as such the inference technique used to learn the mapping must be able to learn a non-linear relational mapping.

A widely adopted model for achieving this is a Bayesian mixture of experts model (BME) [3, 5, 12, 14, 18, 23, 24] where the prediction is formed as a weighted combination of linear models. Each linear model is used to represent a local region of the training set and a logistic regression model is used to give a probabilistic weight to each expert for prediction. This allows the modelling of non-linear mappings through the use of multiple linear models and multi-modality is achieved by different experts representing different modes of the relational mapping. However, typically the pose is inferred by combining the weighted contributions of all experts. In multi-modal regions this has the effect of averaging over the two modes resulting in an incorrect pose.

Another popular technique is Gaussian Processes due to their ability to model non-linear functions, to generalise well from small training sets and accurately model the predictive uncertainty in the mapping [8, 9, 20, 25, 27]. However, due to the $O(N^3)$ computational

 http://dx.doi.org/10.5244/C.26.7

complexity and their unimodal Gaussian predictive distribution, larger and more complex datasets require either significant sparsification [19] or the use of multiple local models [8, 25, 28].

Employing multiple local Gaussian processes [25, 28] overcomes the limitations of Gaussian processes by limiting each models size, thus over coming their $O(N^3)$ complexity and allowing each model to represent a different mode of the mapping. However, previous methods [25, 28] rely on finding nearest neighbours in the feature space making test inference susceptible to noisy image features.

In order to resolve the ambiguities in the multi-modal mapping from image to pose, some researchers have added a dynamical model to select the correct mode by conditioning their pose estimation on the previous frame's pose [23, 28]. These techniques use a first order Markov assumption to learn a dynamical model that is combined with their appearance model to resolve the ambiguities. However, the first order assumption means that the dynamical model also suffers from a high degree of ambiguity. Secondly, these dynamical models are very sensitive to the accuracy of the previous pose estimate [23].

In this paper we propose a mixture of Gaussian Processes model learnt in a similar fashion to a Bayesian mixture of experts model. Our method optimises the size and location of the experts to ensure that they give an accurate predictive distribution over the pose space that is robust to noise in the feature space. To resolve the ambiguities, we introduce a second-order dynamics model that uses dynamic programming to infer an optimum path through our multi-modal predictive distribution. We show that by including a second order term, most of the ambiguity in human pose dynamics is resolved. Our method also alleviates the sensitivity to inaccurate preceding poses by integrating out multiple previous predictions to give a stable pose estimate.

2 Related Work

In this section we give an overview of relevant research in discriminative human pose estimation. The early work on tackling human pose estimation by learning a mapping from image features to the pose space was proposed by Argarwal et al. [1] who used a relevance vector machine to estimate human pose from histogrammed shape context descriptors sampled from silhouettes. Since then a large variety of techniques have been proposed to tackle the problem. As introduced in section 1 Bayesian mixture of experts [3, 6, 12, 14, 18, 23] have received a lot of attention for their simple formulation and ability to model non-linear multi-model relations. Thayananthan et al. [24] adapt the Bayesian mixture of experts by replacing the linear experts with relevance vector machines allowing each expert to model a non-linear function by mapping the inputs through a kernel.

More recently, Gaussian processes have been employed in a variety of ways. Although single Gaussian process models have been employed for pose estimation [27], they are limited to simple data sets due to their uni-modal predictive distribution and $O(N^3)$ complexity. Other approaches use shared space Gaussian process latent variable models (GPLVM) with a dynamical constraint to learn a shared latent space between the image features and the pose space [8, 9, 10, 26]. The dynamical constraint ensures that the latent points are ordered temporally allowing smooth tracking. This model still requires representing the entire training set with a single Gaussian process, restricting its use to small data sets. Urtasun and Darrell [25] construct local online Gaussian process models by selecting expert centres and using neighbouring training points to train small Gaussian process models. Bo and Sminchisescu

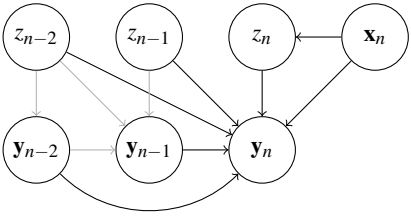

Figure 1: Graphical model for 2nd order pose filtering showing the nodes involved in computing $\mathbf{y}_n$. See section 3.

[5] introduce a twin structured Gaussian process that is able to learn the structure of the pose variables such that appearance predictions correspond to valid poses.

Similar in spirit to [9], Memisevic et al. [17] introduce shared kernel information embedding (sKIE) for pose estimation. This model learns a shared latent space, but in contrast to the GPLVM discussed above, it uses an $O(N^2)$ kernel density estimation to optimise the latent points. They construct local online models centered around the test point to allow inference on large data sets.

The majority of discriminative pose estimation techniques provide an estimation for each image frame independently. Dynamical models play an important role in discriminative tracking to smooth the estimated pose and resolve ambiguities inherent in the image to pose mapping. Sminchisescu et al. [23] introduce a probabilistic framework for discriminative tracking where the predicted pose is conditioned on the previous pose and the image observation. At each frame, they combine the predictions of two BME models, one giving a pose prediction conditioned on the image, the other conditioned on the previous pose. They show that in some sequences, including their pose dynamics is able to reduce tracking errors. However they note that this model is sensitive to the accuracy of previous pose estimates.

Thayananthan et al. [24] use a similar formulation except each Gaussian pose estimation is propagated using a Kalman filter. At each frame they take their propagated density and compute pairwise products with the observation from a mixture of relevance vector machines. Each proposal distribution is then rendered into the image space using a human body model, and they use silhouette chamfer distance to assign each proposal a likelihood. The best K proposal distributions are then propagated to the next frame. To extract a tracked sequence, they use the Viterbi algorithm to find an optimal path through the Kalman filters.

In a similar fashion to a Gaussian process dynamical model, Memisevic et al. [17] propose adding a dynamical constraint to their sKIE model, but they do not provide a full evaluation of this technique.

3 Dynamical Second Order Filtering for Mixture of Experts Models

In this section we introduce an overview of a mixture of experts model and an algorithm that exploits a temporal model to give a smooth prediction from the predictive distribution. A mixture of experts model gives a predictive distribution over the pose $\mathbf{y}$ as a mixture of Gaussian distributions as a function of an image feature $\mathbf{x}$:

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

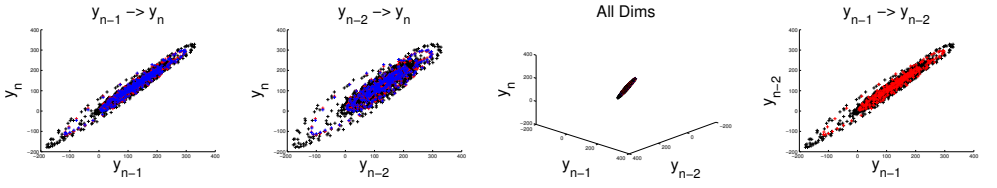

Figure 2: Best viewed in colour and zoomed. These plots show the pose data for the subject's left hand in one axis for the Ballet dataset. Black crosses are training points, red are test and blue are predicted from a linear model $p(y_n|y_{n-1},y_{n-2})$. Plot (a) shows a first order prediction y_{n-1} against y_n, although there is clearly a linear relationship, there is a high degree of ambiguity. Plot (b) shows y_{n-2} against y_n and plot (c) shows a 3D plot with all three variables rotated to demonstrate the linear manifold. The second order pose distribution $p(y_n|y_{n-1},y_{n-2})$ is highly linear, where y_{n-2} resolves vast majority of the ambiguity in plot (a). Plot (d) shows y_{n-1} plotted against y_{n-2}. The prediction of a linear model shown in blue is able to model the human motion to a high degree of accuracy.

Here, 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 $0 \leq p(z=i|\mathbf{x}) \leq 1$.

In human pose tracking we wish to make a point estimate $\hat{\mathbf{y}}_n$ for the pose at frame n. The naive approach is to take the expectation of (1) by taking a weighted average of the component means. While this approach is acceptable in a unimodal setting where only one of the Gaussian components is active, in multi-modal regions this averaging can result in incorrect poses. Secondly, the output of successive frames is often noisy due to the image ambiguities inherent in monocular pose estimation.

We propose an algorithm for inferring a smooth path through sequence of multi-modal pose estimations by forming the problem as a second order Markov model. This allows a dynamics prediction to re-weight the Gaussian components at each frame producing a smooth path through the predictive distribution. By exploiting this property we can use a simple dynamical model which predicts a single mode allowing for exact inference along a Markov chain. This simplifies inference compared to previous methods [23] which used a mixture of experts dynamical model, enforcing the use of a clustering approximation to be made at each frame. We introduce a latent variable to allow for tractable Markov chain inference with the mixture of Gaussians predictive distribution given by the appearance model, see figure 1. The role of this latent variable is to act as a switch, separating out the components of the predictive distribution such that each forms a Gaussian observation. By considering each component in turn, we can obtain an efficient, tractable algorithm for inferring the optimum pose sequence through the Markov chain.

Our model is formulated such that we wish to infer two latent variables at each frame in the sequence, z_n denotes the expert representing a mode of the pose distribution, and $\hat{\mathbf{y}}_n$ represents a smoothed prediction within that mode. That is, we maintain K predictions for each frame in the sequence, we denote the prediction at frame n from appearance expert i as $\hat{\mathbf{y}}_{i,n}$. The predicted pose at frame n by expert $z_n = i$ is given by:

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

where $p(\mathbf{y}_n|\mathbf{x}_n, z_n)$ is the Gaussian expert prediction. $p(\mathbf{y}_n|\hat{\mathbf{y}}_{z_{n-1},n-1}, \hat{\mathbf{y}}_{z_{n-2},n-2})$ is a dynamical prediction which uses the previous two states of $\hat{\mathbf{y}}$ to form a Gaussian prediction for $\hat{\mathbf{y}}_{z_n,n}$. Note the summation of z_{n-1} and z_{n-2}, the dynamical prediction is a weighted sum of

the predictions from all combinations previous locations. This reduces the sensitivity of the prediction to the previous pose estimates. Finally, $p(z_{n-1}|\mathbf{x}_{1:n-1})$ and $p(z_{n-2}|\mathbf{x}_{1:n-2})$ are the marginal probabilities of appearance mode being $z=i$ for a particular frame conditioned on the previous observations. These marginals are evaluated as

$$p(z_n|\mathbf{x}_{1:n}) = p(z_n|\mathbf{x}_n)p(\hat{\mathbf{y}}_{z_n,n}|\mathbf{Y}_{tr}) \sum_{z_{n-2}} \sum_{z_{n-1}} p(z_{n-1}|\mathbf{x}_{1:n-1})p(z_{n-2}|\mathbf{x}_{1:n-2}), \tag{3}$$

where $p(\hat{\mathbf{y}}_{z_n,n}|\mathbf{Y}_{tr})$ is a density model which gives the probability of the posterior pose prediction $\hat{\mathbf{y}}_{z_n,n}$ being a valid pose. This has the effect of encoding the structure between the joints by down weighting pose predictions that aren't globally coherent with the training examples.

The prediction made for each expert at a given frame is a Gaussian product between the dynamical prediction $p(\mathbf{y}_n|\mathbf{y}_{n-1},\mathbf{y}_{n-2})$ and the appearance prediction $p(\mathbf{y}_n|\mathbf{x}_n,\mathbf{z}_n)$. This has the useful property that the influence of each Gaussian distribution is inversely proportional to its uncertainty. That is, if an appearance expert has a higher predictive variance, the dynamics model will have more influence over the prediction, and vice versa. In this paper we model $p(\mathbf{y}_n|\mathbf{y}_{n-1},\mathbf{y}_{n-2})$ using a linear regression model such that

$$p(\mathbf{y}_n|\mathbf{y}_{n-1},\mathbf{y}_{n-2}) = \mathcal{N}(\mu([y_{i,n-1},y_{i,n-2}]^T),\sigma([y_{i,n-1},y_{i,n-2}]^T)) \tag{4}$$

where $\mu()$ and $\sigma()$ are given by standard Bayesian linear regression [4]. We model the dynamics of each joint individually to simplify inference and allow for accurate modelling of the dynamics. Structure between each joint is enforced using a kernel density model $p(\mathbf{y}_*|\mathbf{Y}_{tr}) = \sum_n^N k(\mathbf{y}_*,\mathbf{y}_n)$ where $\mathbf{y}_*$ is the predicted pose and $\mathbf{Y}_{tr} = \{\mathbf{y}_n\}$ are the training poses. Figure 2 shows that a second order linear model is able to give an accurate model of a joint's motion.

Inferring the optimum pose is performed by applying the max-sum algorithm [4] to the sequence. We initialise the algorithm by setting

$$p(\mathbf{y}_1|z_1,\mathbf{x}_1) = \mu_{1,i}, \tag{5}$$
$$p(\mathbf{y}_2|z_2,\mathbf{x}_2) = \mu_{2,i}, \tag{6}$$
$$p(z_1|\mathbf{x}_1) = p(z_1=i|\mathbf{x}_1), \tag{7}$$
$$p(z_2|\mathbf{x}_2) = p(z_2=i|\mathbf{x}_2), \tag{8}$$

where $\mu_{n,i}$ is the predictive mean of component i given by the predictive distribution of appearance model and $p(z_1=i|x_1)$ is the prior associated with each component (equation 1). We then proceed through the sequence evaluating $p(\mathbf{y}_n|z_n,\mathbf{x}_{1:n})$ for each appearance expert z_n and the corresponding marginal probabilities $p(z_n|\mathbf{x}_{1:n})$ (equations 2 and 3). At each frame, we store z_{n-1} and z_{n-2} which correspond to the most likely preceding appearance experts that lead to $z_n=i$:

$$\beta(i,n) = \underset{z_{n-1},z_{n-2}}{\operatorname{argmax}}\, p(z_n=i|\mathbf{x}_n)p(z_{n-1}|\mathbf{x}_{1:n-1})p(z_{n-2}|\mathbf{x}_{1:n-2}) \tag{9}$$

The second phase of the algorithm involves back-tracking through the stored sequences in β to extract the optimum sequence $\mathbf{z}$. We start by setting $z_N = \operatorname{argmax}_i p(z_N=i|x_{1:N})$ and $\hat{\mathbf{y}}_N = p(\mathbf{y}_N|z_N,\mathbf{x}_N)$, and then iterate backwards through the sequence $n=N-1,N-2,\ldots,1$ setting:

$$z_n = \beta(z_{n+1},n+1)_{z_{n-1}}, \tag{10}$$
$$z_{n-1} = \beta(z_{n+1},n+1)_{z_{n-2}}, \tag{11}$$
$$\hat{\mathbf{y}}_n = p(\mathbf{y}_n|z_n,\mathbf{x}_{1:n}). \tag{12}$$

Thus $\hat{\mathbf{Y}} = \{\hat{\mathbf{y}}\}_{n=1}^{N}$ contains the smoothed predicted pose sequence and $\mathbf{z} = \{z_n\}_{n=1}^{N}$ stores the optimal sequence of experts that generated it.

4 Mixtures of Gaussian Process Experts

We construct a mixture of experts model with a Gaussian mixture predictive distribution where each expert is a Gaussian process. Using a Gaussian process for each expert has two distinct advantages compared to previous mixture of experts models [6, 24]. Firstly the predictive uncertainty for a test point, $\sigma_i(\mathbf{x})$ from equation 1, is accurately estimated as a function of the test input. This is particularly valuable when the mixture of experts is employed in the dynamics framework detailed above as the uncertainty of each expert prediction is used to govern its influence in inferring the correct pose.

Secondly a Gaussian process allows the inference of kernel hyper parameters through gradient descent on the likelihood function. This enables the use of a RBF kernel with automatic relevance detection which uses a different length scale for each input dimension. These length scales encode the relative importance of each input feature, making the model more robust to noisy image features [20]. This is a significant advantage compared to other expert models such as a relevance vector machine [6, 24]. These models require the kernel parameters to be learnt using cross-validation, which is impractical for a kernel with automatic relevance detection where there are hundreds of parameters.

The predictions from each expert are combined to give a Gaussian mixture distribution over the pose space. The predictive distribution for a model with K experts is given by

$$p(\mathbf{y}_*|\mathbf{x}_*,\mathbf{X},\mathbf{Y},\theta,\phi,\mathbf{z}) = \sum_{i=1}^{T} p(z_*=i|\mathbf{x}_*,\phi)\,p(\mathbf{y}_*|\mathbf{x}_*,\mathbf{X}_{\vartheta_i},\mathbf{Y}_{\vartheta_i},\theta_i) \tag{13}$$

$$= \sum_{i=1}^{T} p(z_*=i|\mathbf{x}_*,\phi)\,\mathcal{N}(\mu_i(\mathbf{x}_*),\sigma_i(\mathbf{x}_*)). \tag{14}$$

where $\mathbf{z} = \{z_n\}_{n=1}^{N}$, $z_n \in \{1 \ldots K\}$ indicate which expert each data point belongs to, ϑ_i is the set of indices of data points that belong to expert i, $\vartheta_i = \{n : z_n = i\}$. Each prediction is given by a Gaussian process $p(\mathbf{y}_*|\mathbf{x}_*,\mathbf{X}_{\vartheta_i},\mathbf{Y}_{\vartheta_i},\theta)$ trained on a subset of the data ϑ_i and is weighted using a logistic regression model with parameters ϕ that gives the probability of each expert conditioned on the input $p(z_*=i|\mathbf{x}_*,\phi)$.

The expert indicators, $\mathbf{z}$, control the size, location and number of experts and are set by Gibbs sampling over the predictive distribution. The probability of a data point n being assigned to expert i is given by

$$p(z_n=i|\mathbf{z}_{/n},\mathbf{X},\mathbf{Y},\theta_i,\phi) \propto p(\mathbf{y}_n|\mathbf{x}_n,\mathbf{X}_{\vartheta_i/n},\mathbf{Y}_{\vartheta_i/n},\theta_i)\; p(z_n=i|\mathbf{z}_{/n},\mathbf{x}_n,\phi). \tag{15}$$

The distribution $p(z_n=i|\mathbf{z}_{/n},\mathbf{x}_n,\phi)$ gives the probability of the training input $\mathbf{x}_n$ belonging to expert i and is given by an L1 penalized multinomial logistic regression model [4]. L1 penalization results in sparse weights allowing the model to select relevant input features.

This differs from previous models [16, 21] where Dirichlet process is used to create new experts during training. Instead we sample from a multinomial distribution given by $p(z_n=i|\mathbf{z}_{/n},\mathbf{X},\mathbf{Y},\theta_i,\phi)$. Models that place a Dirichlet process over the expert indicators rely on sampling a hyper parameter α from a Gamma distribution $\text{Gamma}(\alpha|a_\alpha,b_\alpha)$. In

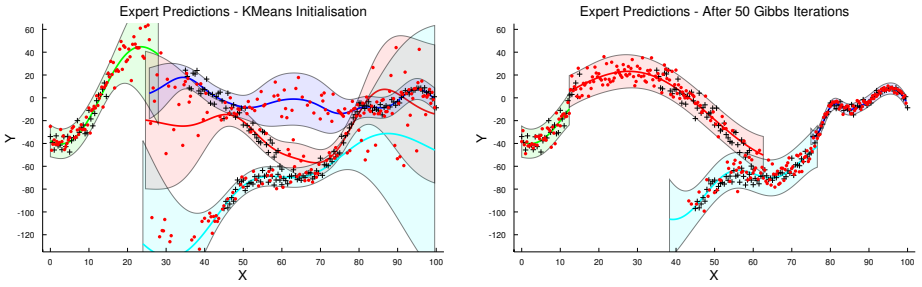

Figure 3: Best viewed in colour. Predictive distributions for the mixtures of Gaussian Processes model on the toy dataset from [16, 21]. The black crosses represent the training points, the red dots are samples drawn from the predictive distribution and the coloured lines represent the predictive mean and variance of each expert. See text for discussion.

practice it is difficult to choose parameters a_α and b_α which lead to a suitable number of experts. Instead, our model is initialised with a large number of small experts, and unsupported experts are removed during the Gibbs sampling process.

The model is formulated in a multivariate setting, such that each expert represents a local set of full poses as opposed to learning a different model for each output dimension. This has the advantage of imposing a degree of structure to the predictive distribution ensuring that predictions made are valid poses as observed from the training set.

When training with large data sets, the size of each expert has to be constrained to avoid individual experts growing such that they are computationally infeasible to train. This is achieved by modifying the distribution given in equation 15 such that the probability of a point being assigned to an expert is zero if it contains a set maximum number of points.

Training is initialised by setting the expert indicators $\mathbf{z}$ either randomly or by running K-means on the training pose data. The algorithm then proceeds in a similar fashion to expectation maximisation. In the expectation step we re-sample the expert indicators and in the maximisation step we update the Gaussian process hyperparameters and the logistic regression weights. To detect convergence we calculate the log likelihood of the training data at each iteration using K-fold cross validation. For a test input $\mathbf{x}_*$, a prediction is made using each expert as in equation 14 and the output is weighted by $p(z_* = i|\mathbf{x}_*, \phi)$.

Figure 3 illustrates the effects of the Gibbs sampling process on a toy data set consisting of 4 functions with varying levels of output noise [16, 21]. The predictive distribution obtained by setting the expert indicators $\mathbf{z}$ using K-means gives poor expert placement resulting in a erroneous predictive distribution. The Gibbs sampling repositions the experts such that each one models a mode of the data set with coherent output noise leading to a much better predictive distribution.

5 Evaluation

We evaluate our method on a Ballet dancing data set [11] and a sign language dataset [2]. We use two types of image features, bag-of-words features [1] and HMAX features [13, 22]. The bag-of-words features have been used extensively in human pose estimation [1, 6, 17, 18]. Our features are computed similarly to [17] using shape context [2] descriptors and 300 cluster centers. For the sign language data set it is not possible to extract silhouettes so we use SIFT descriptors [15]. HMAX features consist of multi-scale Gabor filter responses which are combined using a winner-takes-all max operation to form an image descriptor. We

Dataset:	Ballet			Sign Language	
Model/Feature	BOW SC	HMAX Sil	HMAX	BOW SIFT	HMAX
Our Method (app)	32.52	32.37	32.68	9.50 ± 0.55	6.88 ± 0.48
BME [6]	51.71	61.98	71.72	12.90 ± 0.22	11.87 ± 0.92
Urtasun and Darrell [25]	36.13	33.17	38.18	13.21 ± 1.91	8.11 ± 0.62
sKIE [17]	31.55	31.93	37.57	12.60 ± 0.86	9.36 ± 0.52
Kernel Regression	71.68	71.72	71.71	12.13 ± 0.48	10.65 ± 0.30

Table 1: Quantitive results. Ballet results give the mean average error per joint represented as 3D joint positions in millimeters. Sign language results give the mean absolute error in 2D joint positions in pixels. HMAX Sil and BOW SC features are extracted from silhouettes, HMAX and BOW SIFT are extracted directly from greyscale image.

evaluate these features both with and without silhouettes. Our mixture of Gaussian Processes model is initialised such that each expert contains an average of 100 training points, and are initialised using K-means. Twenty Gibbs sampling iterations are then used to optimise expert locations.

We compare our method against state of the art techniques for monocular pose estimation. The methods we use for comparison are local sKIE [17], Bayesian mixture of linear experts (BME) [6], Urtasun and Darrell's local Gaussian process experts [25] and kernel regression. We set the kernel bandwidth following [17]. All results are calculated by taking the mean absolute error between the predicted pose and the ground truth pose annotations.

We begin by evaluating the appearance model outlined in section 4 against the models outlined above. Following that we compare the effects of applying our dynamical pose filtering (section 3) to enable dynamical tracking.

5.1 Appearance Model

The Ballet data set consists of a complex Ballet dancing sequence performed repeatedly by the same dancer. The choreography is performed 5 times, we use 4 sequences for training and 1 for testing resulting in 1601 frames for training and 356 for testing. Image features are extracted for the whole frame as the dancers movement within the image makes it impractical to extract a tight bounding box around the subject.

Table 1 shows the results for the appearance models. On the silhouette features, we outperform the other mixture of expert models but sKIE has a clear lead. When silhouettes are extracted the inputs are clean enough to allow the isotropic kernel of sKIE to give sufficient accuracy. However, when features are extracted directly from the image, background noise causes the performance of sKIE to degrade.

The Sign Language data set consists of approximately 6000 frames of sign language footage taken from BBC television from a single signer with annotated 2D pose. This is a highly challenging data set due to the complex changing background and rapid motion. We break the sequence into chunks of 400 frames and randomly select a selection of these chunks for training and testing. Our resulting training set consists of 4400 images and the test set has 1200 image. We conduct the experiments on 5 different random partitions of the data. The average results over these 5 partitions are shown in table 1.

Our method has a clear lead over the others on this data set. We suggest that sKIE performs poorly due to the noisy background. Their method doesn't have a mechanism for detecting which are the relevant image features making it highly sensitive to background noise. Urtasun and Darrell [25] method also relies on finding nearest neighbours on the input

Dataset	DPF	LDS	Appearance Only
Sign Language	7.06	7.28	6.88
Ballet	32.19	50.14	32.37

Table 2: Effect of dynamical pose filtering algorithm (DPF) on overall tracking errors.

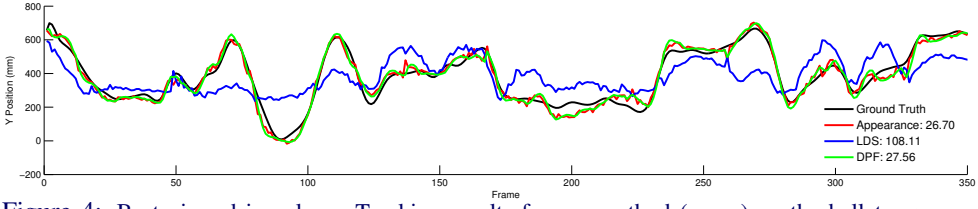

Figure 4: Best viewed in colour. Tracking results for our method (green) on the ballet sequence compared to the appearance model only (red) and a LDS (blue). Plot shows the position of the subject's right hand in the Y axis.

space to build their local Gaussian process models. This makes the model susceptible to input noise, however the Gaussian process experts are able to determine the relevant inputs, leading to an improvement over sKIE. Our model uses a Logistic regression classifier to select the prediction experts which learns a weight for each feature providing robustness to the noisy backgrounds.

5.2 Dynamical Pose Filtering

In this section we evaluate the dynamical pose filtering outlined in section 3 demonstrating it's ability to smooth the predicted pose and resolve ambiguous frames. We compare our model against a standard linear dynamical system and optimise the parameters using expectation maximisation [4]. We manually tuned the dimensionality of the latent space to minimise the tracking errors while remaining numerically stable for the EM algorithm. For the ballet data set the best results were achieved with a 4 latent dimensions, and the sign language used 9. We train the LDS on the ground truth pose data, and then use the Kalman smoothing algorithm [4] to obtain the predicted pose sequence by smoothing $\mathbb{E}[p(\mathbf{y}|\mathbf{x})]$ where $p(\mathbf{y}|\mathbf{x})$ is given in equation 1.

Figure 4 shows an example output of our dynamical pose filtering algorithm compared to the unfiltered expectation of the predictive distribution, $\mathbb{E}[p(\mathbf{y}|\mathbf{x})]$, and a linear dynamical system. Our algorithm smooths the output of the predicted pose while still closely following the predictive distribution of the appearance model. On the other hand, the linear dynamical system makes significant errors and doesn't give a signal that is as smooth.

Table 2 shows the effect on the mean absolute error on both data sets. The dynamical pose filtering does not improve the overall tracking accuracy but gives smoother visual results.

6 Conclusion

In this paper we introduce a tracking system using a novel mixture of Gaussian process experts and a dynamic programming algorithm for inferring a smooth predicted pose for the tracking sequence. We show that this algorithm is competitive with other state of the art techniques and provides a useful system for human pose estimation.

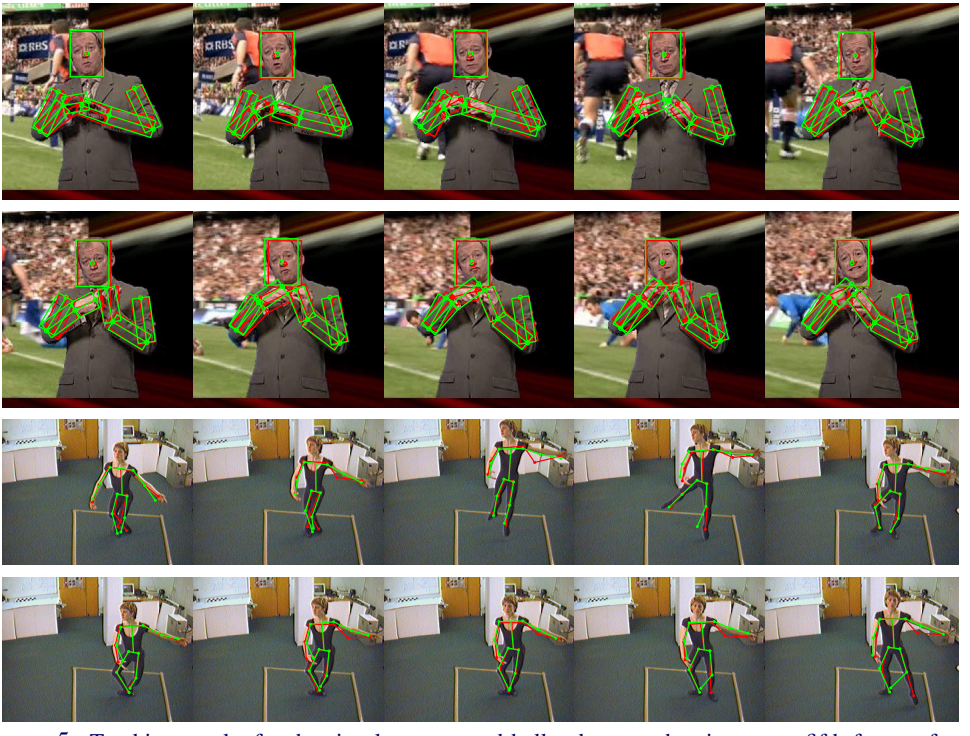

Figure 5: Tracking results for the sign language and ballet datasets showing every fifth frame of a continous sequence. Ground truth shown in red, predicted pose is shown in green.

References

[1] A. Agarwal and B. Triggs. Recovering 3d human pose from monocular images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(1):44–58, Jan. 2006. ISSN 0162-8828. doi: 10.1109/TPAMI.2006.21.

[2] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24 (4):509–522, Apr 2002. ISSN 0162-8828. doi: 10.1109/34.993558.

[3] C. Bishop and M. Svensen. Bayesian hierarchical mixtures of experts. *Uncertainty in Artificial Intelligence*, 2003.

[4] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[5] L. Bo and C. Sminchisescu. Twin Gaussian Processes for Structured Prediction. *International Journal of Computer Vision*, 2010.

[6] L. Bo, C. Sminchisescu, A. Kanaujia, and D. Metaxas. Fast algorithms for large scale conditional 3d prediction. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[7] P. Buehler, MR Everingham, DP Huttenlocher, and A. Zisserman. Long term arm and hand tracking for continuous sign language TV broadcasts. In *Proceedings of the 19th British Machine Vision Conference*, pages 1105–1114. BMVA Press, 2008.

[8] Jixu Chen, Minyoung Kim, Yu Wang, and Qiang Ji. Switching gaussian process dynamic models for simultaneous composite motion tracking and recognition. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 0:2655–2662, 2009. doi: http://doi.ieeecomputersociety.org/10.1109/CVPRW.2009.5206580.

[9] C. H. Ek, P. H. Torr, and N. D. Lawrence. Gaussian process latent variable models for human pose estimation. *A. Popescu-Belis, S. Renals and H. Bourlard (eds) Machine Learning for Multimodal Interaction (MLMI 2007), Springer-Verlag, Brno, Czech Republic*, pages 132–143, 2007.

[10] A. Gupta, T. Chen, F. Chen, D. Kimber, and L.S. Davis. Context and observation driven latent variable model for human pose estimation. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1 –8, 2008. doi: 10.1109/CVPR.2008.4587511.

[11] Shaobo Hou, A. Galata, F. Caillette, N. Thacker, and P. Bromiley. Real-time body tracking using a gaussian process latent variable model. *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8, Oct. 2007. ISSN 1550-5499. doi: 10.1109/ICCV.2007.4408946.

[12] M.I. Jordan and R.A. Jacobs. Hierarchical mixtures of experts and the EM algorithm. *Neural computation*, 6(2):181–214, 1994.

[13] A. Kanaujia, C. Sminchisescu, and D. Metaxas. Semi-supervised hierarchical models for 3d human pose reconstruction. *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pages 1–8, June 2007. doi: 10.1109/CVPR.2007.383341.

[14] Atul Kanaujia and Dimitris Metaxas. Learning ambiguities using bayesian mixture of experts. *Tools with Artificial Intelligence, 2006. ICTAI '06. 18th IEEE International Conference on*, pages 436–440, Nov. 2006. ISSN 1082-3409. doi: 10.1109/ICTAI.2006.73.

[15] D.G. Lowe. Object recognition from local scale-invariant features. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, volume 2, pages 1150–1157 vol.2, 1999. doi: 10.1109/ICCV.1999.790410.

[16] E. Meeds and S. Osindero. An alternative infinite mixture of gaussian process experts. *Advances in Neural Information Processing Systems*, 18:883, 2006.

[17] Roland Memisevic, Leonid Sigal, and David J. Fleet. Shared kernel information embedding for discriminative inference. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(4):778 –790, april 2012. ISSN 0162-8828. doi: 10.1109/TPAMI.2011.154.

[18] Huazhong Ning, Wei Xu, Yihong Gong, and T. Huang. Discriminative learning of visual words for 3d human pose estimation. *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8, June 2008. ISSN 1063-6919. doi: 10.1109/CVPR.2008.4587534.

[19] J. Quiñonero-Candela and C.E. Rasmussen. A unifying view of sparse approximate Gaussian process regression. *The Journal of Machine Learning Research*, 6:1939–1959, 2005.

[20] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005. ISBN 026218253X.

[21] C.E. Rasmussen and Z. Ghahramani. Infinite mixtures of Gaussian process experts. In *Advances in neural information processing systems 14: proceedings of the 2001 conference*, page 881. MIT Press, 2002.

[22] T. Serre, L. Wolf, and T. Poggio. Object recognition with features inspired by visual cortex. *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, 2:994–1000 vol. 2, June 2005. ISSN 1063-6919. doi: 10.1109/CVPR.2005.254.

[23] C. Sminchisescu, A. Kanaujia, Zhiguo Li, and D. Metaxas. Discriminative density propagation for 3d human motion estimation. *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, 1:390–397 vol. 1, June 2005. ISSN 1063-6919. doi: 10.1109/CVPR.2005.132.

[24] A. Thayananthan, R. Navaratnam, B. Stenger, P. Torr, and R. Cipolla. Multivariate relevance vector machines for tracking. *Computer Vision–ECCV 2006*, pages 124–138, 2006.

[25] R. Urtasun and T. Darrell. Sparse probabilistic regression for activity-independent human pose inference. *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8, June 2008. ISSN 1063-6919. doi: 10.1109/CVPR.2008.4587360.

[26] Liang Wang, Xin Geng, C. Leckie, and R. Kotagiri. Moving shape dynamics: A signal processing perspective. *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8, June 2008. ISSN 1063-6919. doi: 10.1109/CVPR.2008.4587554.

[27] Xu Zhao, Huazhong Ning, Yuncai Liu, and T. Huang. Discriminative estimation of 3d human pose using gaussian processes. In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pages 1–4, Dec. 2008. doi: 10.1109/ICPR.2008.4761707.

[28] Xu Zhao, Yun Fu, and Yuncai Liu. Human motion tracking by temporal-spatial local gaussian process experts. *Image Processing, IEEE Transactions on*, 20(4):1141 –1151, april 2011. ISSN 1057-7149. doi: 10.1109/TIP.2010.2076820.