Local Gaussian Processes for Pose Recognition from Noisy Inputs

Martin Fergie
mfergie@cs.man.ac.uk
Aphrodite Galata
a.galata@cs.man.ac.uk

The University of Manchester,
Oxford Road,
Manchester,
M13 9PL, UK

Introduction Gaussian processes have been widely used as a method for inferring the pose of articulated bodies directly from image data. While able to model complex non-linear functions, they are limited due to their inability to model multi-modality caused by ambiguities and varying noise in the data set. For this reason techniques employing mixtures of local Gaussian processes have been proposed to allow multi-modal functions to be predicted accurately [1]. These techniques rely on the calculation of nearest neighbours in the input space to make accurate predictions. However, this becomes a limiting factor when image features are noisy due to changing backgrounds. In this paper we propose a novel method that overcomes this limitation by learning a logistic regression model over the input space to select between the local Gaussian processes. Our proposed method is more robust to a noisy input space than a nearest neighbour approach and provides a better prior over each Gaussian process prediction. Results are demonstrated using synthetic and real data from a sign language data set.

Local Gaussian Process Regression Urtasun and Darrell [1] proposed a model consisting of multiple Gaussian process models each acting as an expert for a region of the data set. Each expert is trained on a neighbourhood of training inputs which are local in pose space. This means that multi-modal regions of the data set will be modelled by placing a Gaussian process expert on each mode. A model with T experts each of size S is trained by selecting T training points as expert centres and learning the hyper-parameters to model its S nearest neighbours in the pose space. The left plot of figure 1 illustrates how the training points of each expert are selected on a toy data set.

Prediction for a test point $\mathbf{x}_*$ is made by selecting the M nearest training points, $\eta_{\mathbf{x}_*} = \{\eta_i\}_{i=1}^{M}$, to the test point $\mathbf{x}_*$ in the input space. A prediction for each η_i is made by selecting the nearest expert to η_i in the pose space, and using S nearest pose neighbours as the training set for that expert. These predictions are then combined as a Gaussian mixture distribution where the distribution over pose is given by

$$p(\mathbf{y}_*|\mathbf{x}_*,\mathbf{X},\mathbf{Y},\Theta) = \sum_{i=1}^{M} \pi_i p(\mathbf{y}_*|\mathbf{x}_*,\mathbf{X}_{\zeta_i},\mathbf{Y}_{\zeta_i},\Theta) = \sum_{i=1}^{M} \pi_i \mathcal{N}(\mu_i(\mathbf{x}_*),\sigma_i(\mathbf{x}_*)), \tag{1}$$

where ζ_i are the nearest neighbours in pose space for η_i and π_i is a function of the inverse prediction variance. The predictive mean and variance, $\mu_i(\mathbf{x}_*)$ and $\sigma_i(\mathbf{x}_*)$, is given by a Gaussian process with covariance $\mathbf{K}_{\zeta_i}$ calculated using training points ζ_i and the hyperparameters from the nearest expert to η_i. Figure 1 shows the training points selected for each expert on the left, and the predictions made for a set of unique test points on the right. The model is able to accurately learn a mapping over the multi-modal region of the data set which a single Gaussian process model would just average.

However, there are a number of limitations with this method. Using the inverse prediction variance, $\sigma_i(\mathbf{x}_*)$, to set the prior over predictions, π_i, results in experts modelling data with noise in the output space to receive a lower weighting. This can be seen in the right panel of figure 1 where the middle region of the data is under represented. Also, the reliance on calculation of nearest neighbours for each test point makes prediction sensitive to noise in the input space. The left panel of figure 2 shows how the prediction accuracy is effected when noise is added to the inputs.

Expert Selection Using Logistic Regression Our proposed method is inspired by a mixtures of experts model and uses a logistic regression model to select between the Gaussian process experts in an equivalent role to π_i in equation 1. Logistic regression models the probability of an input $\mathbf{x}$ belonging to a class c_i. The probability of the class conditioned on the input, $p(c_i|\mathbf{x})$ is given by

$$p(c_i|\mathbf{x}) = \frac{e^{(\mathbf{w}_i^T\mathbf{x})}}{\sum_{j=1}^{T} e^{(\mathbf{w}_j^T\mathbf{x})}}. \tag{2}$$

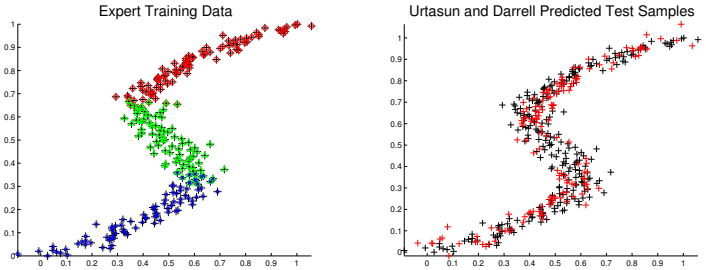

Figure 1: Local Gaussian process experts; training data is shown in black. The left plot shows the points selected for training each Gaussian process expert in green, red and blue. The right plot shows in red the predicted values y_* for an independent test set.

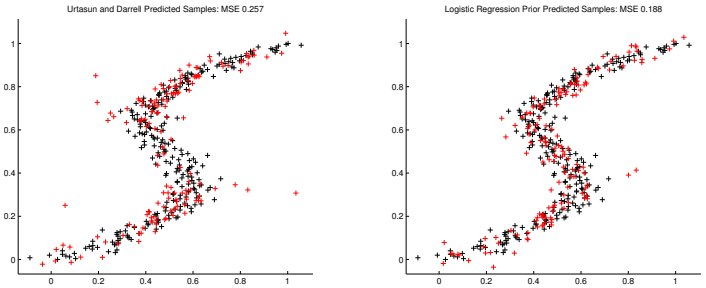

Figure 2: Toy problem as in figure 1 with noise added to the input space. The left panel shows results using [1]. The noise has caused the predictions to be chosen from incorrect experts resulting in lower prediction accuracy. The right panel shows prediction using the method outlined in this paper.

This is a linear combination of the inputs $\mathbf{x}$ and a vector of weights $\mathbf{w}_i$ which is then evaluated with the softmax function such that $p(c_i|\mathbf{x}) \in (0,1)$.

Training the local Gaussian process regression model is similar to [1] except the extra step of training the logistic regression model. The logistic regression model is trained by assigning a class label c_i to each training point (x_n, y_n) which indicates which expert that training point is closest to in pose space. Predictions for a test point $\mathbf{x}_*$ are now made by making a prediction from all individual experts and using $p(c_i|\mathbf{x}_*)$ as a prior on the predictions, replacing π_i in equation 1. The predictive distribution now becomes

$$p(\mathbf{y}_*|\mathbf{x}_*,\mathbf{X},\mathbf{Y},\Theta) = \sum_{i=1}^{T} p(c_i|\mathbf{x}_*)\mathcal{N}(\mu_i(\mathbf{x}_*),\sigma_i(\mathbf{x}_*)), \tag{3}$$

where $p(c_i|\mathbf{x}_*)$ is given by the logistic regression model, and T is the number of experts. The predicted mean and variance for each expert no longer make predictions using the local neighbourhood ζ_i as no nearest neighbours are calculated. Instead, the training set that was used to train each expert ϑ_i is used. Figure 2 shows how this method alleviates the issues caused by noisy inputs compared to [1].

Experimental Evaluation The method has been evaluated on a data set of an actor performing sign language with manually annotated 2D joint positions. The below table shows the mean absolute error of our method compared to [1]. Prediction using our method is roughly 8 times faster as covariance matrices are not recomputed at each test point.

Image Feature	Urtasun and Darrell [1]	Logistic Regression
HMAX	9.09	8.73
SIFT Grid	10.67	8.90

[1] R. Urtasun and T. Darrell. Sparse probabilistic regression for activity-independent human pose inference. *CVPR 2008*, pages 1–8, June 2008.