Semi-supervised Learning of Joint Density Models for Human Pose Estimation

Ramanan Navaratnam * Andrew Fitzgibbon † Roberto Cipolla *

* University of Cambridge
Department of Engineering
Cambridge, CB2 1PZ, UK
http://mi.eng.cam.ac.uk/∼cipolla

† Microsoft Research Ltd
7 JJ Thomson Avenue
Cambridge, CB3 0FB, UK
http://research.microsoft.com/∼awf

Abstract

Learning regression models (for example for body pose estimation, or BPE) currently requires large numbers of training examples—pairs of the form (image, pose parameters). These examples are difficult to obtain for many problems, demanding considerable effort in manual labelling. However it is easy to obtain unlabelled examples—in BPE, simply by collecting many images, and by sampling many poses using motion capture. We show how the use of unlabelled examples can improve the performance of such estimators, making better use of the difficult-to-obtain training examples.

Because the distribution of parameters conditioned on a given image is often multimodal, conventional regression models must be extended to allow for multiple modes. Such extensions have to date had a pre-set number of modes, independent of the contents of the input image, and amount to fitting several regressors simultaneously. Our framework models instead the joint distribution of images and poses, so the conditional estimates are inherently multimodal, and the number of modes is a function of the joint-space complexity, rather than of the maximum number of output modes.

We demonstrate the improvements obtainable by using unlabelled samples on synthetic examples and on a real pose estimation problem, and demonstrate in both cases the additional accuracy provided by the use of unlabelled data.

1 Introduction

We are interested in estimating the parameters of complex parametrized models from a single image, for example the 3D position and joint angles of the human body. Given an image[1] $\mathbf{x}$, and denoting by θ the vector of unknown parameters, we wish to compute $p(\theta|\mathbf{x})$, the probability density over the parameters θ conditioned on the image. This density will in general be multimodal, and we will learn the mapping from $\mathbf{x}$ to $p(\theta|\mathbf{x})$ from training examples. Traditionally this training data is a set of pairs $\{(\mathbf{x}_i, \theta_i)\}_{i=1}^{D}$ where D is the number of training pairs. In many situations, however, training data is expensive to obtain, making it difficult to learn models of any complexity. The contribution of this paper is to introduce a framework which allows additional *unlabelled* examples, of

[1] Throughout this document we shall denote images by real vectors. These vectors may be thought of as vectors of raw image pixels, or—as is used in our experiments—the bins of a shape-context histogram [3, 5, 13].

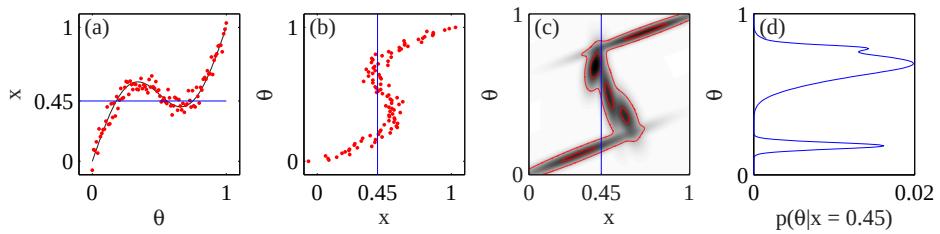


Figure 1: **Generative, discriminative, and joint estimation**. (a) Samples from the generative model $x = f(\theta) + \varepsilon$ where $f(\theta) = \theta + 0.3\sin(2\pi\theta)$ and $\varepsilon \sim \mathcal{N}(0, 0.05)$. Given a value of x, represented by the horizontal line, the corresponding parameters θ must be discovered by search. (b) The inverse relationship is multivalued: for each x there may be more than a single θ. (c) Modelling the joint density $p(x, \theta)$ allows the conditional $p(\theta|x = 0.45)$ to be obtained representing the multiple modes as shown in (d). In this paper we address the problem of learning the joint density from impoverished training sets.

the form $\{(*, \theta_j)\}_{j=1}^{D_\theta}$ and $\{(\mathbf{x}_k, *,)\}_{k=1}^{D_x}$, to contribute to the estimate. This extends two strands of recent computer vision and machine learning research: regression-based body pose estimation and semi-supervised regression.

1.1 Regression-based body pose estimation

Recent approaches to the estimation of model parameters from images divide into two schools: *generative* and *discriminative*. Generative approaches make the assumption that an observed image is drawn from density which is unimodal given the parameters θ. For example an image may be generated by the function $\mathbf{f}(\theta)$ and observed with added noise ε drawn from an isotropic Gaussian:

$$\mathbf{x} = \mathbf{f}(\theta) + \varepsilon, \qquad \varepsilon \sim \mathcal{N}\left(\mathbf{0}, \sigma^2 \mathtt{I}\right) \tag{1}$$

In order to obtain an estimate of θ for a given image $\mathbf{x}$, generative approaches perform a search in θ space, for example using Newton iterations (e.g. extended Kalman filter tracking [8, 9]), or particle filtering [14]. Because of the need for initial estimates of θ, generative techniques tend to imply a tracking framework, where the parameters computed from the previous image serve as an initial estimate for that in the current.

Discriminative, or regression-based, techniques[2], on the other hand, may be considered as trying to learn the distribution over θ directly from the image $\mathbf{x}$, via a function $\mathbf{g}(\mathbf{x})$. When dealing with simple models, this amounts to learning the inverse mapping $\theta = \mathbf{f}^{-1}(\mathbf{x})$, so that model parameters are computed from the input image without recourse to search. Rosales *et al*. [10] and Agarwal and Triggs [2] applied this technique to body pose estimation, while Williams *et al*. [18] applied a similar strategy to face tracking.

In general, however, there will be ambiguities because a given image could have resulted from any of several parameter values. For example, the same human silhouette can

[2]This is a slight variation on the use of "discriminative" in classification problems, but is current in the computer vision literature.

Figure 2: **Pose ambiguities:** This illustrates ambiguities arising from silhouettes. The same human silhouette can be generated from many body positions.

be generated from many body positions (see Fig 2). Thus the mapping must be multivalued, in general returning the parameters ϕ of a distribution in θ space, with ϕ a function of the image $\mathbf{x}$. For example, ϕ may contain the parameters of a M-Gaussian mixture model (GMM), so that

$$\phi(\mathbf{x}) = [\alpha_{1..M}(\mathbf{x}), \mu_{1..M}(\mathbf{x}), \Sigma_{1..M}(\mathbf{x})]$$

and the distribution over θ is then

$$p(\theta|\mathbf{x}) = \sum_{m=1}^{M} \alpha_m(\mathbf{x}) \mathcal{N}(\theta; \mu_m(\mathbf{x}), \Sigma_m(\mathbf{x}))$$

where the dependence of the GMM parameters ϕ on $\mathbf{x}$ has been made explicit. Thayananthan *et al*. [15] modelled this mapping using a multivariate extension of the relevance vector machine framework [16], where $\mu_m(\mathbf{x})$ is a radial basis function gated by $\alpha_m(\mathbf{x})$ and Σ_m is diagonal and independent of $\mathbf{x}$. In their case, $\alpha_m(\mathbf{x})$ is obtained through likelihood evaluation by doing model projections at each $\mu_m(\mathbf{x})$.

Sminchisescu *et al*. [13] also propose a discriminative framework where multinomial regressors are used to model the gating functions ($\alpha_m(\mathbf{x})$) whose parameters are obtained from the training set. The problem is that this requires a lot of training data for each value of $\mathbf{x}$, or equivalently, that there is little generalization from one silhouette to another. Both of the aforementioned methods however do not utilise the marginal data.

1.2 Joint density modelling

More recently, Agarwal [3] proposed learning the *joint* density $p(\mathbf{x}, \theta)$, in order to allow generalization across input examples. By fitting a GMM to the training pairs, we obtain an expression for the joint density of the form

$$p(\mathbf{x}, \theta) = \sum_{m=1}^{M} \alpha_m \mathcal{N}\left(\begin{bmatrix} \mathbf{x} \\ \theta \end{bmatrix}; \begin{bmatrix} \mu_m^x \\ \mu_m^\theta \end{bmatrix}, \begin{bmatrix} \Sigma_m^{xx} & \Sigma_m^{x\theta} \\ \Sigma_m^{\theta x} & \Sigma_m^{\theta\theta} \end{bmatrix} \right) \tag{2}$$

However, in their formulation, the covariance matrices are of a restricted form constructed from the regression error and the input-space distribution. Given an observed image $\mathbf{x}$, the density over θ is given by the standard GMM conditional [4]:

$$p(\theta|\mathbf{x}) = \sum_{m=1}^{M} \alpha_m \mathcal{N}\left(\theta; \mu_m^\theta + \Sigma_m^{\theta x}(\Sigma_m^{xx})^{-1}(\mathbf{x} - \mu_m^x),\ \Sigma_m^{\theta\theta} - \Sigma_m^{\theta x}(\Sigma_m^{xx})^{-1}\Sigma_m^{x\theta}\right) \tag{3}$$

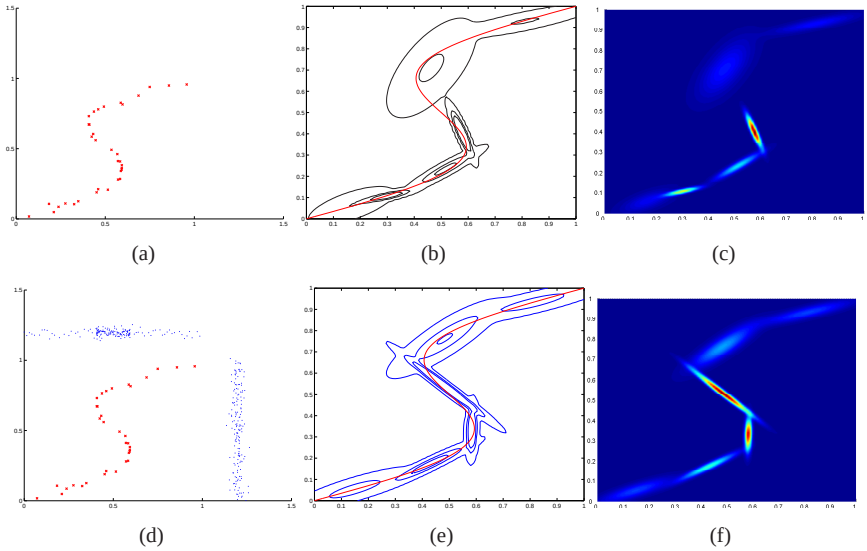


Figure 3: **Learning is improved by the use of unlabelled data.** (a) 35 samples from the joint distribution. Contours (b) and height map (c) of the fitted 5-GMM. The contours are superimposed on the true function (red curve). (d) 35 samples from the joint distribution, augmented with 165 samples from each of the marginals (in blue). (e,f) The fitted GMM incorporating marginal data. Note that even this small amount of marginal data considerably improves the representation of the underlying function.

Figure 1 illustrates this approach on a toy example, showing how the GMM yields a multimodal distribution for $p(\theta|\mathbf{x})$. However fitting the joint-space density still requires a large amount of labelled data. In the next sections we will show how this requirement can be reduced by the use of unlabelled samples.

1.3 Semi-supervised learning

Semi-supervised learning began as a means of improving classifier systems. A classifier is conventionally trained by supplying (image, label) pairs, denoted $\{(\mathbf{x}_i, \theta_i)\}_{i=1}^{D}$ in our notation, where the labels θ come from a small discrete set. The task of classification is to learn a function f such that $f(\mathbf{x}_i) = \theta_i$ for as many of the training examples as possible. Semi-supervised classification adds additional examples of the form $\{(\mathbf{x}_i, *)\}_{j=1}^{D_x}$, without any labels. The additional examples help generalization because they describe the space in which the $\mathbf{x}$'s lie, allowing labels to propagate from labelled areas of the space to unlabelled areas via the prior density $p(\mathbf{x})$. Without the additional examples, classifers must make strong assumptions on the form of $p(\mathbf{x})$, for example approximating it by a Euclidean-metric Parzen window estimate. Semi-supervised learning has been shown to provide improvements in applications such as object recognition [17], digit classification [20] as well as on the UCI classification database [12]. Semi-supervised regression [19] extends the application domain to the case where θ may be a real function, but does not deal with multivalued outputs (indeed it does not return a distribution over θ), and

does not make use of the output-space distribution $p(\theta)$, as we do in this paper. Roth *et al*. [11] proposed a method to learnt the joint-space distribution from marginal data for constructing better likelihood models from image feature measurements. However, they used marginal samples only and did not use any joint-space data.

2 Semi-supervised joint density modelling

In this paper we show how re-casting the joint density modelling approach as a missing data problem allows the use of unlabelled data. At the same time we extend semi-supervised regression to return a distribution over the output space.

We are given three types of training data: associated (image, pose) pairs $\{(\mathbf{x}_i, \theta_i)\}_{i=1}^{D}$; images without poses $\{(\mathbf{x}_k, *)\}_{k=1}^{D_x}$; and poses without images $\{(*, \theta_j)\}_{j=1}^{D_\theta}$. Our task is to learn the parameters ϕ of a Gaussian mixture model which maximizes the likelihood of the training data. We proceed by writing the complete data log likelihood including the unknown variables $\mathbf{z}_{im}$ that indicate i^{th} data point is generated by m^{th} mixture component. The likelihood is optimized using the method of Ghahramani and Jordan [7], and is described below for completeness.

Let $\mathbf{u}_i$ denote the data pair $(\mathbf{x}_i, \theta_i)$. Let the superscripts $*$ and o indicate subvectors and submatrices of the parameters ϕ matching the missing and observed components of the data respectively.

$$\begin{aligned}\ln(P(\phi|\mathbf{u},\mathbf{z})) = \sum_{i=1}^{D+D_x+D_\theta}\sum_{m=1}^{M}\mathbf{z}_{im}\Big[&\frac{d}{2}\ln 2\pi + \frac{1}{2}\ln|\Sigma_m| - \frac{1}{2}(\mathbf{u}_i^o-\mu_m^o)^T(\Sigma_m^{oo})^{-1}(\mathbf{u}_i^o-\mu_m^o)\\ &-(\mathbf{u}_i^o-\mu_m^o)^T(\Sigma_m^{o*})^{-1}(\mathbf{u}_i^*-\mu_m^*) - \frac{1}{2}(\mathbf{u}_i^*-\mu_m^*)^T(\Sigma_m^{**})^{-1}(\mathbf{u}_i^*-\mu_m^*)\Big] \quad (4)\end{aligned}$$

As in the standard EM algorithm [6] used to learn the parameters of a GMM, consider the expectation of the equation 4. One would require the unknown terms $\mathbf{z}_{im}$, $\mathbf{z}_{im}\mathbf{u}^*$ and $\mathbf{z}_{im}\mathbf{u}^*\mathbf{u}^{*T}$ to obtain the sufficient statistics for the parameters ϕ. These can be defined as follows.

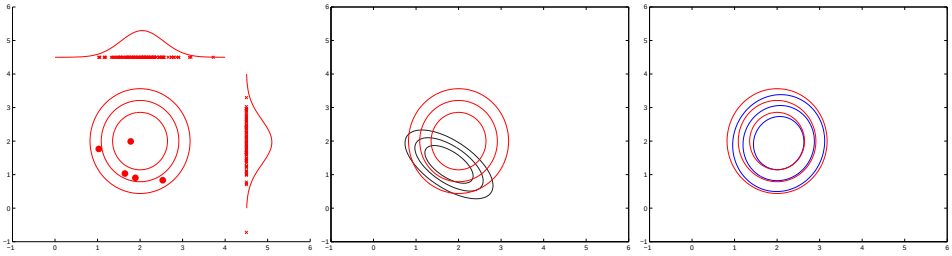
$$E[\mathbf{z}_{im}|\mathbf{u}_i^o,\phi_{m|t-1}] = \frac{\mathcal{N}\left(\mathbf{u}_i^o : \phi_{m|t-1}^o\right)}{\sum_{m=1}^{M}\mathcal{N}\left(\mathbf{u}_i^o;\phi_{m|t-1}^o\right)} \quad (5)$$

$$E[\mathbf{u}_i^*|m,\mathbf{u}_i^o,\phi_{m|t-1}] = \mu_{m|t-1}^* + \Sigma_{m|t-1}^{*o}(\Sigma_{m|t-1}^{oo})^{-1}(\mathbf{u}_i^o-\mu_{m|t-1}^o) \quad (6)$$

$$= \mathbf{u}_{i|t}^* \quad (7)$$

$$E[\mathbf{u}_i^*\mathbf{u}_i^{*T}|m,\mathbf{u}_i^o,\phi_{m|t-1}] = \Sigma_{m|t-1}^{**} - \Sigma_{m|t-1}^{*o}(\Sigma_{m|t-1}^{oo})^{-1}\Sigma_{m|t-1}^{*oT} + \mathbf{u}_{i|t}^*\mathbf{u}_{i|t}^{*T} \quad (8)$$

These are substituted appropriately in the standard EM update equations for learning the parameters ϕ of a GMM. In the equation 8, the terms $\Sigma_{m|t-1}^{**} - \Sigma_{m|t-1}^{*o}(\Sigma_{m|t-1}^{oo})^{-1}\Sigma_{m|t-1}^{*oT}$ quantify the uncertainty associated with the estimated missing values. This is a key difference from a naive update that uses only $\mathbf{u}_{i|t}^*\mathbf{u}_{i|t}^{*T}$ for the covariance update.



(a) (b) (c)

Figure 4: **Contribution of unlabelled data.**This illustrates how marginal data are used to enhance the estimate of a 2D Gaussian using only 5 joint-space data and 100 marginal sample in each dimension. Figure 4(a) shows the true Gaussian from which the data is drawn. It also shows the joint-space and marginal data. Figure 4(b) superimposes the Gaussian learnt using just the 5 joint-space points by explicitly computing the mean and covariance (black contours) on the true distribution (red contours). Figure 4(c) shows the Gaussian computed using the marginal as well as the joint-space data (blue contours) along with the true distribution (red contours).

2.1 Why and how can unlabelled data help?

The reader may wonder why the above algorithm yields a better fit to the underlying distribution than simply using the labelled pairs. The results we show later, and the examples in figure 3 confirm that there is indeed a practical advantage, but it is instructive to consider on an intuitive level what the marginal samples can contribute to the estimation of the joint density. First, consider the estimation of a 2D Gaussian distribution using "labelled" points (x_i, y_i). We must determine five parameters: the mean (μ_x, μ_y) and the upper triangle of the covariance matrix $(\sigma_{xx}, \sigma_{xy}, \sigma_{yy})$ [Fig. 4(b)]. Now assume we also have "unlabelled" points x_j and y_k, and for convenience, assume we have so many unlabelled points that we in fact know the marginal distributions $p(x)$ and $p(y)$ perfectly [Fig. 4(c)]. For a Gaussian, this already exactly constrains the mean, and two degrees of freedom in the covariance matrix. Thus the number of parameters we need to estimate is reduced from five to one. Intuitively, we now require five times fewer labelled examples to have the same quality of estimate. Of course we do not have perfect knowledge of the marginals when we have a finite number of unlabelled examples, but the data contribute in the same way to estimation of the joint density.

3 Results

We performed a number of experiments to test the method. First was a simple quantitative test on the toy data of Figure 1. Firstly a large number (500) of joints-space samples were created and a 6 mixture GMM was learnt using these to represent the "ground-truth" distribution.

Then 6-mixture GMMs were trained using varying number of joint-space and marginal samples. The fit to the ground-truth distribution was measured as the total negative log likelihood of 500 test data points. We expect that for small numbers of joint-space samples, this fit will be poor. The question then is to what extent adding marginal samples

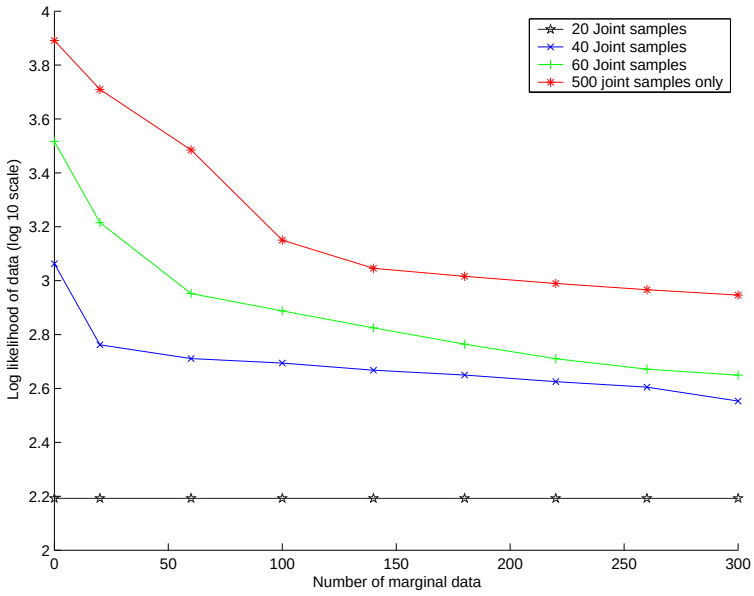


Figure 5: **Contribution from marginal data.**This illustrates how much the marginal data are contributing towards obtaining a better estimate of the underlying distribution. For example, 20 joint samples plus 300 marginals gives as good a fit as using 60 joint samples only. The multi-modal function $\mathbf{x} = \theta + 0.3\sin(\theta) + \mathcal{N}(0, 0.05)$, was used to create the data.

will improve the estimate. Figure 5 shows the results for various numbers of samples. For example, for 20 samples in the joint space, the initial fit is very poor, but adding 300 marginal samples brings it close to the ground truth value. Reading the graph another way, 20 joint samples plus 300 marginals gives as good a fit as using 60 joint samples only. Given that marginal samples are essentially free to obtain, this is a cost reduction of a factor of 3 in obtaining training data.

We performed experiments on a body pose estimation problem demonstrating the applicability of the method in a practical example. The input image descriptor $\mathbf{x}$ is constructed from shape contexts of silhouette points. To work in a common coordinate system, these features are clustered into 40 groups (in our experiments) and each feature point is projected onto the common basis by weighted voting into the cluster centres [13, 3]. All these feature vectors for a silhouette are added to create a feature histogram to represent the image descriptor ($\mathbf{x}$) for that silhouette. The output vector θ is the first 6 (in our experiments) principal components of the 56-dimensional joint angle space. Though we focused on walking motion in our experiments, no action specific constraints were used.

GMMs were trained with varying number of mixture components as well as joint-space samples and marginal samples. For comparison, additional GMMs were learnt using the joint samples alone to highlight the improvement gained by exploiting the marginal data. All samples were obtained using real motion captured data [1] and a custom built graphical package that renders images from these motion data in order to allow reference against the ground truth. We tested the GMMs on 1000 synthetic images created

by the graphical package and 60 hand-labelled real images. The optimum number of mixture components were identified through cross-validation. Results of these experiments are provided in Table 1. Comparing the RMS errors either in column 3 or in column 4, the results indicate that marginal data indeed contribute towards learning a better joint-space distribution.

Joint-space data	Marginal data	RMS on artificial images	RMS on real images
2 000	-	6.2	11.0
2 000	8 000	5.9	10.6
1 000	9 000	6.2	11.9
10 000	-	5.6	10.0

Table 1: **Quantitative experiments:**
Results of the quantitative experiments are tabulated above. In these experiments, 1000 artificial images (created using a graphical package) and 60 hand-labelled images were used to compute the average RMS error in the estimated pose. Comparing the RMS errors either in column 3 or in column 4, the results indicate that marginal data indeed contribute towards learning a better joint-space distribution.

Some estimated poses along with their likelihood and RMS error in the predicted joint angle for a few of the real images were shown in figure 6. The prediction with the best likelihood did not always produce the estimate with the least RMS error. However, the important point to note is that all the predictions were possible hypothesis for the input silhouette.

4 Discussion

Regression models for pose estimation based on variants of the mixture of experts paradigm [3, 12, 14] have exploited labelled training examples only. In this paper, we show that using cheaply available marginal, or unlabelled, data along with the labelled samples enhances regression models. The additional improvement obtained by a GMM based regressor has been demonstrated clearly in synthetic examples and a real world body pose estimation problem. We emphasize in this paper that not only the joint-space distribution, but also the marginal distributions can be harnessed to improve regression models by exploiting unlabelled data. Moreover, the number of multiple hypotheses predicted by regressors based on mixture of experts models is less than or equal to the number of mixtures. However, regressors based on joint-space distribution are not constrained as such. The point to note is that the choice of the number of mixture components has little effect on the number of hypothesis predicted for an 'ambiguous' input, i.e. the choice of M is not the same as choice of arity.

As future work, we intend to investigate an interactive learning strategy based on the fact that a small number of joint-space examples and a large amount of marginal data are enough to capture the joint-space distribution reasonably well for regression. The initial joint-space is obtained from marginals only. Afterwards, the user adds informative/ intelligent joint-space samples to guide the regressors to achieve better accuracy. Hence the required accuracy can be achieved with the minimum number of joint-space samples, thus reducing the time consuming task of labelling.

Acknowledgments

This work was supported by the Gates Cambridge Trust, the ORS Programme, and Toshiba Research. The authors wish to thank Oliver Williams and Arasanathan Thayananthan for numerous discussions, and the anonymous reviewers for their helpful comments.

References

[1] http://mocap.cs.cmu.edu/.

[2] A. Agarwal and B. Triggs. 3d human pose from silhouettes by relevant vector regression. In *CVPR*, 2004.

[3] A. Agarwal and B. Triggs. Monocular human motion capture with a mixture of regressors. In *IEEE Workshop on Vision for HCI at CVPR*, 2005.

[4] T. W. Anderson. *An Introduction to Multivariate Statistical Analysis*. John Wiley & Sons, Inc., New York, 3rd edition, 2003.

[5] S. Belongie, J. Malik, and J. Puzicha. Matching shapes. In *ICCV*, volume I, pages 454–461, 2001.

[6] A.P. Dempster, N. Laird, , and D.B. Rubin. Maximum likelihood from incomplete data via the em algorithm. In *J. of the Royal Statistical Society, Series B*, volume 34, pages 1–38, 1977.

[7] Zoubin Ghahramani and Michael I. Jordan. Supervised learning from incomplete data via an EM approach. In *Proc. NIPS*, volume 6, pages 120–127, 1994.

[8] N. Jojic, M. Turk, and T.S. Huang. Tracking self-occluding articulated objects in dense disparity maps. In *ICCV*, pages 123–130, 1999.

[9] K. Rohr. Towards model-based recognition of human movements in image sequences. *CVGIP: Image Understanding*, 59(1):94–115, June 1994.

[10] R. Rosales and S. Sclaroff. Inferring body pose without tracking body parts. In *CVPR*, pages II: 721–727, 2000.

[11] S. Roth, L. Sigal, and M.J. Black. Gibbs likelihoods for bayesian tracking. *CVPR*, 1:886–893, 2004.

[12] N. Shental, T. Hertz, A. Bar-Hillel, and D. Daphna Weinshall. Computing gaussian mixture models with em using side-information. 2003.

[13] C. Sminchisescu, A. Kanaujia, Z. Li, and D. Metaxas. Discriminative density propagation for 3d human motion estimation. 2005.

[14] J. Sullivan and J. Rittscher. Guiding random particles by deterministic search. In *ICCV*, volume I, pages 323–330, 2001.

[15] A. Thayananthan, R. Navaratnam, B. Stenger, P. H. S. Torr, and R. Cipolla. Multivariate relevance vector machines for tracking. In *ECCV*, 2006.

[16] Michael E. Tipping. Sparse bayesian learning and the relevance vector machine. *J. Mach. Learn. Res.*, 1:211–244, 2001.

[17] J. J. Verbeek and N. Vlassis. Semi-supervised learning with gaussian fields technical report. Technical Report IAS-UVA-05-01, University of Amsterdam, The Netherlands, 2005.

[18] O. Williams, A. Blake, and R. Cipolla. A sparse probabilistic learning algorithm for real-time tracking. In *ICCV*, pages 353–360, 2003.

[19] Z. Zhou and M. Li. Semi-supervised regression with co-training. In *IJCAI*, 2005.

[20] X. Zhu and Z. Ghahramani. Learning from labeled and unlabeled data with label propagation. Tech. report in preparation, Carnegie Mellon University, 2002.

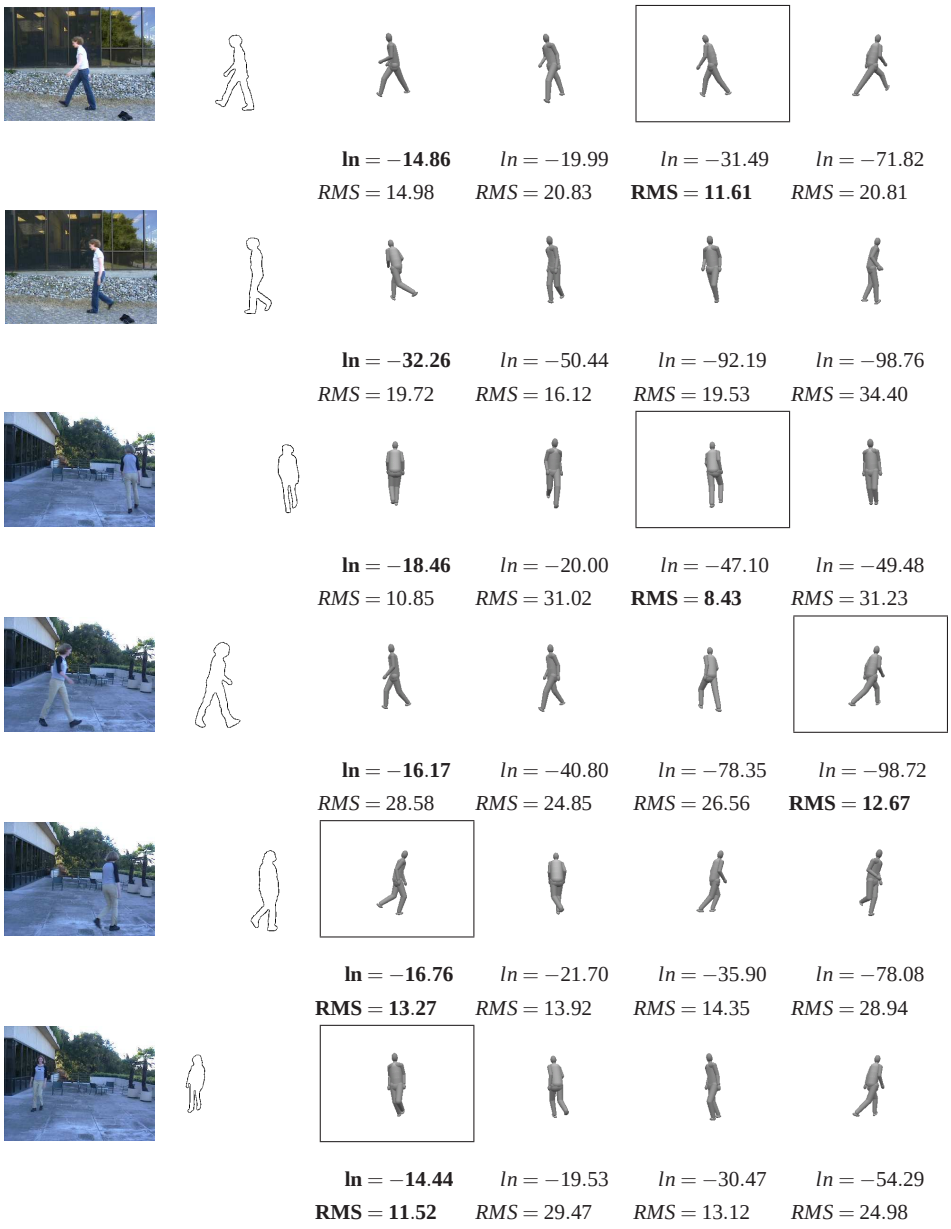


Figure 6: **Pose Detection:** *This illustrates results from applying the GMM learnt from 8k marginal and 2k joint data points with 50 mixture components. First column shows the real images. The second column shows the silhouette edge image. The final four columns show the best four modes ordered based on the log likelihood, which are also shown below respective images. In the second row, the first four estimates do not contain the correct pose as the human in the image.*