Estimating 3D Pose via Stochastic Search and Expectation Maximization

Ben Daubney and Xianghua Xie

Department of Computer Science, Swansea University,
United Kingdom, SA2 8PP
{B.Daubney,X.Xie}@swansea.ac.uk

Abstract. In this paper an approach is described to estimate 3D pose using a part based stochastic method. A proposed representation of the human body is explored defined over joints that employs full conditional models learnt between connected joints. This representation is compared against a popular alternative defined over parts using approximated limb conditionals. It is shown that using full limb conditionals results in a model that is far more representative of the original training data. Furthermore, it is demonstrated that Expectation Maximization is suitable for estimating 3D pose and better convergence is achieved when using full limb conditionals. To demonstrate the efficacy of the proposed method it is applied to the domain of 3D pose estimation using a single monocular image. Quantitative results are provided using the HumanEva dataset which confirm that the proposed method outperforms that of the competing part based model. In this work just a single model is learnt to represent all actions contained in the dataset which is applied to all subjects viewed from differing angles.

Key words: 3D Pose Estimation, Expectation Maximization, Stochastic Search, Rigid Joint, Loose Limbed

1 Introduction

There is currently much interest in being able to extract the pose of a human from a single or sequence of images. A popular technique used to achieve this is to represent the human body as a probabilistic graphical model, where the nodes of the graph represent anatomical parts of the body and the edges represent the relationships between these parts [1–4, 12–14]. However, a limitation with current part based methods is the use of the Loose Limbed model, which approximates the joint between two connected parts using a soft connection. This representation does not enforce the connecting joint between neighboring parts to coincide and is employed as the likelihood of two neighboring parts being detected independently, with their connecting joints exactly aligned, is very low. In this work a method is presented that uses ancestral sampling to generate a set of hypothesis locations where the connecting joint between neighboring parts is constrained to coincide.

Forcing joints between connected parts to coincide will address one the key limitations with the current Loose Limbed approach and will result in a model

that is better constrained and is a more intuitive representation of the human body constructed of rigid parts with fixed joint locations. To achieve this, rather than defining a model over parts/limbs as is usual in current Loose Limbed approaches [1–4, 12, 14], we define a model where the hidden nodes of the graph represent joint locations. This proposed representation is referred to as a Fixed Joint model.

A further limitation with current Loose Limbed approaches is that typically the conditional probability distribution used to represent the relationship between neighboring parts, referred to as a limb conditional in this work for brevity, is approximated by learning a distribution over the relative state between connected limbs [1–3, 14]. This is motivated by our knowledge of the human body; a given joint has a fixed and known range over which it can move. However, in order to learn approximate limb conditionals the original training data must be converted into a relative form. This process eliminates much of the original data's structure, therefore any model learnt using this will fail to capture its full complexity. In this work it is shown that learning a full conditional model between connected parts provides a richer and much more accurate description of the training set and therefore the object being modeled.

The principal reason that human pose estimation is difficult is the large number of degrees of freedom that the human body contains. Attempts to efficiently search this space using a graphical part based representation of the human body include Dynamic Programming [1] and Belief Propagation [4, 13] for 2D pose estimation and stochastic methods such as the Pampas algorithm [6], Variational MAP [7] and Partitioned Sampling [8] for 3D pose estimation. These methods are iterative and require that a model must first be defined to propagate the particle set between iterations of the algorithm; how this model is defined is not intuitive and often the covariance of this model is simply initially overestimated and then shrunk at each iteration to force convergence [7, 9]. A motivation for using full limb conditionals is that pose can then be efficiently estimated using Expectation Maximization (EM) and importance sampling. At each iteration samples are drawn from the prior which are then weighted to approximate the posterior distribution given the current observations, using this sample set the prior is then reestimated. In the following iteration a new set of samples are drawn from the reestimated prior and over a number of iterations the prior converges to a solution; empirically this solution appears to be global. Using this method samples are always drawn from the prior and an extra model to propagate samples between iterations is not necessary. A further advantage of this approach is that it results in a compact parametric description of the posterior distribution. This parametric representation is particularly advantageous in applications such as tracking where drift between frames could be added deterministically by scaling the resultant covariances.

In this work three principal claims are made: Firstly, compared to the Loose Limbed representation the proposed Fixed Joint model results in a prior that is far more representative of the original training set. Secondly, that the proposed Fixed Joint model results in faster convergence of the EM algorithm compared

to the Loose Limbed model. Thirdly, that the Fixed Joint model outperforms that of the Loose Limbed model at estimating pose. These claims are supported by both quantitative and qualitative results using the HumanEva data set [5]. Whilst the presented approach is general enough that it could readily be applied to scenes captured from multiple views or employed in a tracking framework, here it is applied to single images and it is assumed that the position of the root node is fixed and known *a priori*. We employ this constrained scenario as the focus of this paper is on highlighting the limitations of existing representations and demonstrating the advantages of the proposed method through detailed analysis and comparison of performances. This is best achieved by constraining any experiments so that observed differences in performance can only be a direct result of the methodology used. However, the presented approach is adequately efficient such that uncertainty in the root node could be accommodated by sampling the root position multiple times, however, this is currently left for future work.

2 Pose Estimation

The problem of estimating pose of an articulated object can be defined over a probabilistic graph where the set of n hidden nodes $v_i \in \mathcal{V}$ represent the set of parts used to represent the object and $\{v_i, v_j\} \in \mathcal{E}$ represent the edges that connect the nodes of the graph. Given a set of proposal values for each node $X = \{\mathbf{x}_i, .., \mathbf{x}_n\}$ and a set of observations for each node $Z = \{z_i, .., z_n\}$ the posterior can then be calculated as

$$p(X|Z, \theta) = \prod_{\{i,j\} \in \mathcal{E}} p(\mathbf{x}_i|\mathbf{x}_j, \theta_{ij}) \prod_{i \in \mathcal{V}} p(z_i|\mathbf{x}_i) \tag{1}$$

where $\mathbf{x}_i$ is assumed to be the child of $\mathbf{x}_j$, $p(\mathbf{x}_i|\mathbf{x}_j, \theta_{ij})$ are limb conditionals which represent the model prior and θ_{ij} is a connection parameter, and $p(z_i|\mathbf{x}_i)$ are observational likelihoods. Pose can then be estimated by finding the configuration X^* that maximizes this equation. It is assumed that the graph used to represent the articulated object is a tree and therefore contains no loops.

The focus of this paper is on the comparison between using a Loose Limbed model defined over parts and a proposed Rigid Joint model defined over joint positions. As discussed in the proceeding section, whilst the Loose Limbed model approximates the limb conditional $p(\mathbf{x}_i|\mathbf{x}_j, \theta_{ij})$ from Equation 1 with a model learnt over $\mathbf{x}_i$ in the frame of reference of $\mathbf{x}_j$ denoted by $p(\mathbf{x}_{ij}|\theta_{ij})$, the Rigid Joint model uses full limb conditionals $p(\mathbf{x}_i|\mathbf{x}_j, \theta_{ij})$ which we show to be both far more representative of the original training set and result in faster convergence of the EM algorithm. In the following Sections we describe the limb conditionals learnt for each model, how samples can be generated from these models and how Equation 1 is maximized using EM.

2.1 Model Representation

Loose Limbed Model The Loose Limbed model is based on that presented in [2] which we briefly describe. The model is defined over parts and each part has 6 degrees of freedom $\mathbf{x}_i = (\mathbf{r}_i, \mathbf{\Theta}_i)$, where $\mathbf{r}_i \in \mathbf{R}^3$ and $\mathbf{\Theta}_i \in \mathbf{SO}(3)$ which

represent the global position of the proximal joint of the ith part and its rotation respectively, each part has a fixed length. The rotations are represented by unit quaternions, therefore $\mathbf{x}_i \in \mathbf{R}^7$. Rather than learning a conditional distribution over $\mathbf{x}_i$ and $\mathbf{x}_j$ directly a distribution is instead learnt over $\mathbf{x}_{ij}$, where $\mathbf{x}_{ij}$ is the position and orientation of the ith part described in the local frame of reference of the jth part. Given a set of training data the distribution $p(\mathbf{x}_{ij}|\theta_{ij})$ can be learnt directly for each part using a GMM. Following [2] each limb conditional is represented using three components.

Rigid Joint Model The proposed Rigid Joint model is defined over joint positions, where the distance between neighboring joints is fixed. Conditional models $p(\mathbf{x}_i|\mathbf{x}_j, \theta_{ij})$ are learnt where $\mathbf{x}_i$ is the orientation of the ith joint defined in a global frame of reference (i.e. that of the root node). These models are also learnt using a GMM.

To create a conditional model a joint distribution $p(\mathbf{x}_i, \mathbf{x}_j|\theta_{ij})$ is first learnt from which the conditional distribution can be calculated during run time as described in Section 2.2. A prior distribution over the position of each joint is learnt over spherical coordinates (ρ, θ, ϕ), where ρ represents the length between the joint and the joint to which it is connected, $\theta \in [0, 2\pi]$ represents a rotation around the xy-plane and $\phi \in [0, \pi]$ represents the elevation measured relative to the z-axis. Since the length is fixed ρ is constant for each joint and we have only two free parameters θ and ϕ, which describe the orientation of each joint measured in the global frame of reference (i.e. that of the root node). We represent these two angles using polar coordinates (r, ω), where the rotation $\omega = \theta$ and the radius $r = \phi$, where $r \in [0, \pi]$.

The limitation with this representation is that a discontinuity occurs at $r = \pi$. To overcome this we also create a duplicate polar coordinate system where $r = \pi - \phi$ so that at the origin $\phi = \pi$. Each coordinate system is referred to using the suffixes 0 and π respectively as this indicates the value ϕ at the origin. Each position in the coordinate system also has a weight associated with it such that those nearer the origin are weighted higher that those near the outer edges (i.e. near the discontinuity) these weights are defined as $w_0 = \frac{r}{\pi}$ and as $w_\pi = 1 - w_0$. Hence, a measurement represented in 3D spherical coordinates $\mathbf{x} = (\rho, \theta, \phi)$ is thus represented as a set of 2D vectors and weights $\mathbf{x} = \{\mathbf{x}_0, w_0, \mathbf{x}_\pi, w_\pi\}$, where $\mathbf{x}_0 = (r_0, \omega_0)$. Using this representation a GMM could be learnt for each coordinate system independently and weighted proportional to the total weight of the training data used. These weights then describe whether the data was distributed near to the origin of the coordinate system, where it is better represented, or near the edge, where the discontinuity occurs and it is poorly represented.

Given training data for two connected joints i and j, $\mathbf{X}_i = \{[\mathbf{x}_i]_1, .., [\mathbf{x}_i]_l\}$ and $\mathbf{X}_j = \{[\mathbf{x}_j]_1, .., [\mathbf{x}_j]_l\}$, where $\mathbf{x}_i = \{\mathbf{x}_0^i, w_0^i, \mathbf{x}_\pi^i, w_\pi^i\}$ and l is the number of samples in the training set, a joint distribution is learnt by first concatenating the two sets of training data together so that $\mathbf{X}_{ij} = \{[\mathbf{x}_{ij}]_1, .., [\mathbf{x}_{ij}]_l\}$ where $\mathbf{x}_{ij} = (\mathbf{x}_i, \mathbf{x}_j)$. Using this data the joint probability distribution $p(\mathbf{x}_i, \mathbf{x}_j|\theta_{ij})$ can be estimated, however, as each training point is represented by a set of two

vectors and two weights, $\mathbf{x}_i = \{\mathbf{x}_0^i, w_0^i, \mathbf{x}_\pi^i, w_\pi^i\}$ and $\mathbf{x}_j = \{\mathbf{x}_0^j, w_0^j, \mathbf{x}_\pi^j, w_\pi^j\}$, when concatenating the data we must do so for each possible combination of the ordering of ϕ, i.e. $\mathbf{x}_{ij} = \{\mathbf{x}_{00}^{ij}, \mathbf{x}_{0\pi}^{ij}, \mathbf{x}_{\pi 0}^{ij}, \mathbf{x}_{\pi\pi}^{ij}\}$ where for example $\mathbf{x}_{0\pi}^{ij} = \{(\mathbf{x}_0^i, \mathbf{x}_\pi^j), w_{0\pi}^{ij}\}$ and the corresponding scalar weights are simply multiplied together so that $w_{0\pi}^{ij} = w_0^i w_\pi^j$. The consequence of this is that for each pair of connected joints we have four sets of training data, a GMM is learnt for each independently. Each GMM is assigned a weight proportional to the total weight of the training set (e.g. $W_{0\pi}^{ij} = \sum_{k=1}^{l} \left[w_{0\pi}^{ij}\right]_k$) so that GMM's with more data clustered near the origin have a higher weight since these will better represent the data. The prior of each individual GMM component is then scaled by this weight.

The number of components used to represent each distribution in the model is set to reflect the increasing complexity in the distribution at nodes located at a further depth from the root node. To represent this we employ the following scheme: Joints connected directly to the root node are given three components and at every subsequent increase in depth a further two components are added. Under our model the maximum number of components is assigned to the wrists with nine components. Whilst this may immediately seem advantageous since the Rigid Joint model is afforded a maximum of nine components compared to the Loose Limbed's three, it should be noted that the rigid model's distribution must represent a far larger space; it is likely the three component distribution of the Loose Limbed model is far more representative of the training data once it has been converted into a relative form. Our argument is that in the process of converting the original training set so that a Loose Limbed model can be learnt a large amount of information is being discarded from it.

2.2 Sampling

As the graphical model used to represent the articulated object is a tree and the root node is assumed to be fixed, samples can be generated using ancestral sampling [10]. Samples are drawn hierarchically starting from those nodes closest to the root node, then at each step down the tree, moving away from the root node, a further set of samples can be drawn conditioned on those samples generated for the parent node. To efficiently search the pose space the number of particles are exponentially grown moving out from the root node. This ensures that the location of less constrained joints are searched using more samples. For each sample $\mathbf{x}_j^m$, N child samples are drawn from the limb conditional $[\mathbf{x}_i^n]_{n=1}^N \sim p(\mathbf{x}_i|\mathbf{x}_j^m, \theta_{ij})$, N is referred to as the growth rate. As very few particles are needed to describe the prior distribution for nodes near the root this exponential growth is not problematic, for example setting $N = 8$ will result in 4096 samples being generated for each of the wrists. For efficiency all covariances used to represent limb conditionals are assumed to be diagonal.

Loose Limbed Model As the Loose Limbed model only uses an approximated limb conditional a sample $\mathbf{x}_i^n$ can be generated from a parent sample $\mathbf{x}_j^m$ simply by drawing a sample $\mathbf{x}_{ij}^n \sim p(\mathbf{x}_{ij}|\theta_{ij})$, which can then be transformed into the

global frame of reference through $M(\mathbf{x}_i^n) = M(\mathbf{x}_{ij}^n)M(\mathbf{x}_j^m)$, where $M(\mathbf{x}_i^n)$ represents the 3D object-to-world transform. To draw a sample from this distribution a GMM component k^* is sampled from the marginal distribution $p(m_{ij}^k) = \lambda_{ij}^k$, where the connection parameters $m_{ij}^k = \{\mu_{ij}^k, \Sigma_{ij}^k, \lambda_{ij}^k\}$ define the mean, covariance and weighting of the kth component of the GMM respectively, following which a sample for $\mathbf{x}_{ij}^n$ can be drawn from $\mathbf{x}_{ij}^n \sim \mathcal{N}(\mu_{ij}^{k^*}, \Sigma_{ij}^{k^*})$.

Rigid Joint Model Given a sample for the jth node $\mathbf{x}_j^m$, a sample can be drawn conditioned on this by first calculating the marginal likelihood of observing this for each component in the GMM. Given that all covariance matrices are diagonal, i.e. $\Sigma_{ij}^k = diag(\Lambda_{ii}^k, \Lambda_{jj}^k)$, the marginal likelihood is given by $p(\mathbf{x}_j^m|m_{ij}^k) = \lambda_{ij}^k \mathcal{N}(\mathbf{x}_j^m; \mu_j^k, \Lambda_{jj}^k)$. Once this has been calculated for all components the resultant distribution is normalized to give the conditional distribution $p(m_{ij}^k|\mathbf{x}_j^m)$. A GMM component can then be sampled from this distribution $k^* \sim p(m_{ij}^k|\mathbf{x}_j^m)$, from which a sample $\mathbf{x}_i^n$ can be drawn from the selected component $\mathbf{x}_i^n \sim \mathcal{N}(\mu_i^{k^*}, \Lambda_{ii}^{k^*})$. Notice that in the case of the Loose Limbed model $p(m_{ij}^k|\mathbf{x}_j^m) = p(m_{ij}^k)$ i.e. is independent of $\mathbf{x}_j^m$.

2.3 Rigid Joint Model: Observing a Joint

The problem in defining a model over joints as apposed to parts is that there does not exist one-to-one correspondences between joints and observations; we can not directly observe a joint only the parts to which it is connected. To accommodate this we define a set of m observable parts $p_i \in P$, where $m \neq n$ and n represents the number of joints in the model. We further define $v_j \in p_i$ as being the set of joints defining the ith part and conversely $p_j \in v_i$ as being the set of parts of which the ith joint is a member. The set of observations made for the parts are defined by $Z = \{\mathbf{z}_i, .., \mathbf{z}_m\}$. The observational likelihood for the ith part can be written as $p(\mathbf{z}_i|\{\mathbf{x}_{j\in p_i}\})$, where this distribution is dependent on a number of joint positions. Intuitively, this represents that for example the appearance of the forearm must be dependent on the location of both the wrist and elbow. To estimate $p(\mathbf{z}_i|\mathbf{x}_j)$ from $p(\mathbf{z}_i|\{\mathbf{x}_j, \mathbf{x}_{k\in p_i|j}\})$ the nodes $\mathbf{x}_{k\in p_i|j}$ can be treated as nuisance parameters and marginalized over. In practice this is cumbersome to calculate and instead the following approximations are used: If the $\mathbf{x}_{k\in p_i|j}$ are child nodes to $\mathbf{x}_j$ we calculate $p(\mathbf{z}_i|\mathbf{x}_j)$ using the expectation of the set of particles drawn from $\mathbf{x}_j$ as $\mathbf{x}_{k\in p_i|j}$. If they are parent nodes we use the sample of $\mathbf{x}_{k\in p_i|j}$ from which $\mathbf{x}_j$ was drawn. For the torso we use the expectation of the shoulder and hips since these joints are not directly connected and do not share child/parent relationships. This method then allows an approximation of the term $p(\mathbf{z}_i|\mathbf{x}_j)$ to be calculated.

We further need to account for that a joint may be a member of several parts, for example the elbow defines both the upper arm and forearm. To accommodate this the likelihood terms $p(\mathbf{z}_i|\mathbf{x}_j)$ are combined for all parts to which that joint is a member $p_i \in v_j$. This can be calculated as

$$p(\mathbf{z}_{i\in v_j}|\mathbf{x}_j) = \prod_{i\in v_j} p(\mathbf{z}_i|\mathbf{x}_j). \tag{2}$$

This suggests that to infer the position of a joint all parts to which it is connected must be observed. Whilst in this section we have described how the observation likelihood is calculated for a joint we will write $p(\mathbf{z}_{i\in v_j}|\mathbf{x}_j)$ as $p(\mathbf{z}_j|\mathbf{x}_j)$ so that the same notation can be used when describing optimization of both the Loose Limbed and Rigid Joint model in the following section.

2.4 Maximization

Maximizing the posterior is achieved using EM where a new prior is estimated at each iteration given the posterior calculated using the old prior (M-step), a new set of particles is then generated from the prior and the posterior re-estimated (E-step). Given a set of M particles for the jth joint $\left[\mathbf{x}_j^m\right]_{m=1}^M$ each is assigned a weight proportional to the marginal likelihood $p(\mathbf{x}_j^m|\mathbf{z}_j)$. This can efficiently be calculated for each node using a simplified form of the Sum-Product algorithm. The outwards messages from the root node are represented by the generated set of samples and as such only backwards messages must be computed. Due to the ancestral sampling method used this can be efficiently calculated, the marginal for sample $\mathbf{x}_j^m$ is computed as

$$p(\mathbf{x}_j^m|\mathbf{z}_j) = p(\mathbf{z}_j|\mathbf{x}_j^m) \prod_{i\in C_j} \sum_{n=1}^{N} p(\mathbf{x}_i^n|\mathbf{z}_i) \tag{3}$$

where $i \in C_j$ is the set of nodes that are the children of the jth node and the summation is performed over the set of N samples that were drawn conditioned on the sample $\mathbf{x}_j^m$ under the ancestral sampling method.

At each iteration simulated annealing is used to ensure the distribution converges so that $w_j^m = p(\mathbf{x}_j^m|\mathbf{z}_j)^\beta$, where β is calculated at each iteration so approximately 60% of the particles would be discarded if resampling were performed [9]. Given the set of weighted samples the prior can then be reestimated.

2.5 Limb Likelihoods

A part is represented by a rectangular patch and defined by the joints that it is composed from (Rigid Joint) or the proximal/distal joint of the part (Loose Limbed). We use two image cues, edges and color. Edge cues are exploited using a set of M overlapping HOG features [11] placed along the edges of the part. Each feature is represented as a single normalized histogram of the local image gradients and are combined such that $p(\mathbf{z}_j|\mathbf{x}_j)_{edge} = \frac{1}{M}\prod_{m=1}^{M} H(\theta_\perp)$, where $H(\theta_\perp)$ returns the value in the histogram bin that is perpendicular to the edge of the proposed part.

Color is exploited by placing a bounding box at the location of the root node and then learning a foreground model using the pixel values within the box and a model for the background using pixels outside the box. The models are learnt using a GMM. This creates a very crude and noisy foreground probability map, the likelihood is then calculated as the average foreground probability value encompassed by the part. The individual likelihoods for each cue are then combined as $p(\mathbf{z}_j|\mathbf{x}_j) = p(\mathbf{z}_j|\mathbf{x}_j)_{edge}p(\mathbf{z}_j|\mathbf{x}_j)_{col}$.

3 Experiments

Both a Rigid Joint and Loose Limbed model were learnt using the Train partition of the HumanEva dataset using ≈ 4500 frames of data taken across all subjects and actions. Samples drawn from the prior of each model can be seen in Fig. 1 along with the training data from which the models were learnt. It is clear in this figure that the samples drawn from the Rigid Joint model much more closely resemble that of the training data, the samples from the Loose Limbed model are much more broad and shows less clear structure, this is particularly clear on the feet.

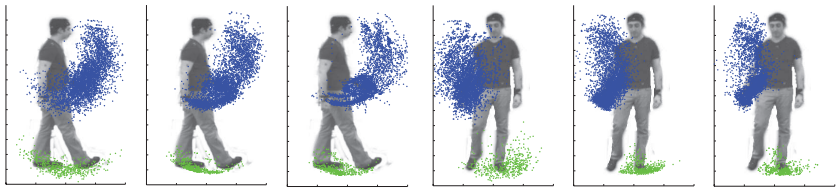

Fig. 1. Comparing samples of the left foot (green) and right wrist (blue) generated by each model representation and the training data. Side View: (a) Loose Limbed model (b) Rigid Joint model (c) Training data. Frontal view: (d) Loose Limbed model (e) Rigid Joint Model (f) Training data.

To compare the performance of both models a test set was created from the Validation partition of the HumanEva dataset. This was composed of 100 randomly selected frames from each action category (Box, Gesture, ThrowCatch, Walk, Jog) selected across all color views and all subjects, so that 500 frames were used in total. The root node and orientation was set using the pelvis marker data from the groundtruth provided and the scale was set as the maximum distance between the head and the feet. This scale is often inaccurate (e.g. if the subject was squatting) however, is used so all experiments are easily reproducible.

Both methods used the same settings so that the only difference in each experiment was the model used. Quantitative results can be see in Fig. 2 where it can be seen that the Rigid Joint representation outperforms the Loose Limbed model. We also experimented between updating the model by calculating marginals using Equation 3 or simply using local image evidence (i.e. setting $p(\mathbf{x}_j^m|\mathbf{z}_j) = p(\mathbf{z}_j|\mathbf{x}_j^m)$). As shown the use of marginals improves the error, this is because these allow information about observations being made at the extremities of the tree to influence the convergence of those parts nearer the root node.

In Fig. 3 the expected pose and samples drawn from the prior are presented after each iteration for the example shown, as can be seen the Rigid Joint model converges much faster than the Loose Limbed model. Notice also the slip between the parts of the lower left leg in Fig. 3 (a) (v) this is as joint positions are not constrained to coincide in the Loose-Limbed representation.

To illustrate why a conditional model converges more efficiently than an approximated conditional model consider Fig. 4, which shows a hypothetical multimodal distribution. Whilst the full limb conditional model can converge, the relative limb conditional can not until its parent's limb conditional has converged to a single mode. In Fig. 5 an example is shown using a growth rate $N = 2$, this

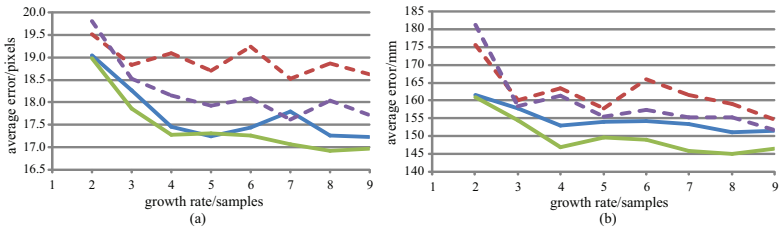

Fig. 2. Pose estimation errors as a function of growth rate (N) for 2D (a) and 3D (b) pose estimation after ten iterations of the algorithm. Dashed lines represent Loose Limbed model and solid lines Rigid Joint model. The green and purple line show the error using full marginals and the blue and red line shows the error using only local image evidence.

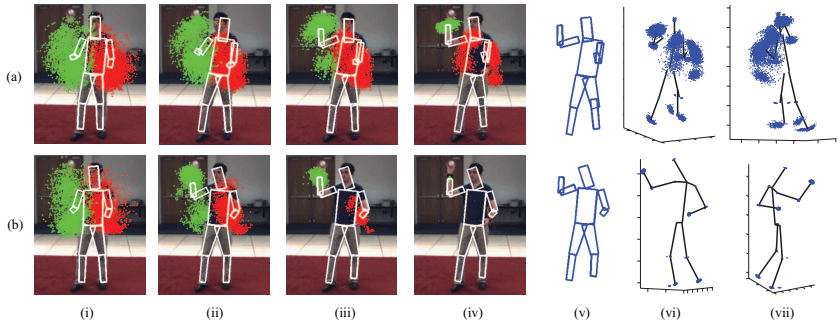

Fig. 3. Example of convergence for Loose Limbed model (a) and Rigid Joint model (b). (i) to (iv) shows iterations 1, 3, 5, 10 respectively. Samples for the left (red) and right (green) wrist drawn from each prior are also shown as is the expected pose. (v) shows the final expected pose. (vi) and (vii) show the final 3D reconstruction with samples that have been drawn from the final model.

uses just a maximum of 16 samples for the wrists. However, as can be seen the presented method still finds the correct solution, it is the performance using very few samples that is particularly impressive and makes this approach of value.

4 Conclusions

A method has been presented to estimate 3D pose from single images using a stochastic search and Expectation Maximization. A novel part based representation has been defined over joint positions and compared against an existing method, it has been shown quantitatively that the presented method outperforms that of the Loose Limbed model. Furthermore, we have demonstrated qualitatively that using full limb conditionals results in a model that is more representative of the original training set and efficiently converges under the EM algorithm. Whilst in this paper it has been assumed the root node is fixed the approach can be generalized to account for uncertainty in this value by sampling multiple root node positions and will be the focus of future work.

References

1. Felzenswalb, P.F., Huttenlocher, D.P.: Pictorial Structures for Object Recogntion. International Journal on Computer Vision, vol. 61, pp. 55-79 (2005)

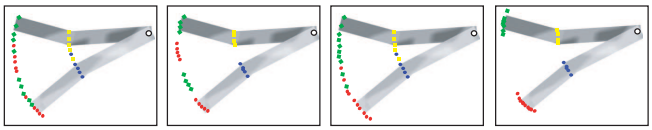

Fig. 4. Hypothetical two part example highlighting the difference in convergence between a relative limb conditional (a) and (b) and a full limb conditional (c) and (d). (a) and (c) show the prior model and (b) and (d) the model after a number of iterations. Both limb conditionals are represented by a two component GMM where each component is represented by different colors. Whilst the conditional model can represent each observational mode by a single Gaussian (d), the relative model can not and as such 'phantom modes' appear in the prior (b) slowing convergence.

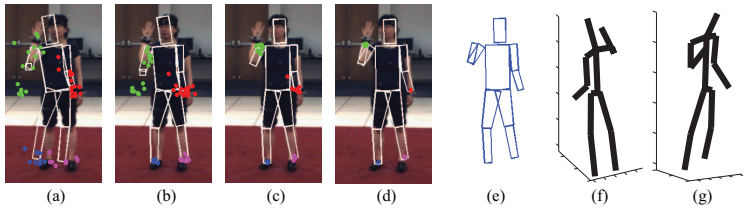

Fig. 5. Example of convergence and 3D pose estimation using a growth rate $N = 2$. (a) - (d) Iteration 1, 2, 4 and 10 respectively. (e) Expected pose as shown in (d). (f) and (g) final 3D expected pose.

2. Sigal, L., Bhatia, S., Roth, S., Black, M.J., Isard, M.:Tracking Loose-Limbed People. CVPR, pp. 421-428 (2005)
3. Andriluka, M., Roth, S., Schiele, B.: Pictorial Structures Revisited: People Detection and Articualted Pose Estimation. CVPR, pp. 1-8 (2009)
4. Ramanan, D.: Learning to Parse Images of Articulated Bodies. NIPS, pp. 1129-1136 (2006)
5. Sigal, L., Balan, A., Black, M.: HumanEva: Synchronized Video and Motion Capture Dataset and Baseline Algorithm for Evaluation of Articulated Human Motion. International Journal of Computer Vision, vol. 87, pp. 4-27 (2009)
6. Isard, M.: PAMPAS: Real-Valued Graphical Models for Computer Vision. CVPR, pp. 613-620 (2003)
7. Hua, G., Wu, Y.: Variational Maximum a Posteriori by Annealed Mean Field Analysis. IEEE Transaction on Pattern Analysis and Machine Intelligience, vol. 27, pp. 1747-1761 (2005)
8. Deutscher, J., Davidson, A., Reid, I.: Automatic Partitioning of High Dimensional Search Space associated with Articulated Body Motion Capture. CVPR, pp. 669-676 (2001)
9. Deutscher, J., Blake, A., Reid, I.: Articualted Body Motion Capture by Annealed Particle Filtering. CVPR, pp. 126-133 (2000)
10. Bishop, C., M.: Pattern Recognition and Machine Learning. Springer (2006)
11. Dalal, N., Triggs, B.: Histogram of Orientated Gradients for Human Detection. CVPR, pp. 886-893 (2005)
12. Ferrari, V., Marin-Jimenez, M., Zisserman, A.: Progressive Search Space Reduction for Human Pose Estimation. CVPR, pp. 1-8 (2008)
13. Gao, J., Shi, J.: Multiple Frame Motion Inference using Belief Propagation. IEEE Conference on Automatic Face and Gesture Recognition, pp. 875-880 (2004)
14. Bernier, O., Cheung-Mon-Chan, P.: Real-Time 3D Articualted Pose Tracking using Particle filter and Belief Propagation on Factor Graphs. BMVC, pp. 27-36 (2006)