Predicting 3D People from 2D Pictures

Leonid Sigal Michael J. Black

Department of Computer Science, Brown University, Providence, RI 02912
{ls,black}@cs.brown.edu

Abstract. *We propose a hierarchical process for inferring the 3D pose of a person from monocular images. First we infer a learned view-based 2D body model from a single image using non-parametric belief propagation. This approach integrates information from bottom-up body-part proposal processes and deals with self-occlusion to compute distributions over limb poses. Then, we exploit a learned Mixture of Experts model to infer a distribution of 3D poses conditioned on 2D poses. This approach is more general than recent work on inferring 3D pose directly from silhouettes since the 2D body model provides a richer representation that includes the 2D joint angles and the poses of limbs that may be unobserved in the silhouette. We demonstrate the method in a laboratory setting where we evaluate the accuracy of the 3D poses against ground truth data. We also estimate 3D body pose in a monocular image sequence. The resulting 3D estimates are sufficiently accurate to serve as proposals for the Bayesian inference of 3D human motion over time.*

1 Introduction

The estimation of 3D human pose and motion is relatively well understood in controlled laboratory settings with multiple cameras where any number of Bayesian inference methods can recover 3D human motion (e.g. [4]). All of these methods rely on accurate background subtraction and edge information; this is a strong limitation that prevents their use in more realistic and complex environments. When the background is changing or the camera is moving, reliable background subtraction is difficult to achieve. The problems become particularly acute in the case of monocular tracking where the mapping from the 2D image to the 3D body model is highly ambiguous. Solutions to the monocular (static camera) case have relied on strong prior models [18], manual initialization [23] and/or accurate silhouettes [1, 2, 19, 23]. The fully automatic case involving a monocular camera is the focus of this paper.

Recent work on 2D body pose estimation and tracking treats the body as a "cardboard person" [9] in which the limbs are represented by 2D planar (or affine) patches connected by joints. Such models are lower-dimensional than the full 3D model and recent work has shown that they can be estimated from 2D images [5, 14, 15]. The results are typically noisy and imprecise but they provide exactly the kind of information necessary to generate *proposals* for the probabilistic inference of 3D human pose. Thus we simplify the 3D problem by introducing an intermediate 2D estimation stage.

Acknowledgments. This work was partially supported by Intel Corporation and NSF IGERT award #9870676.

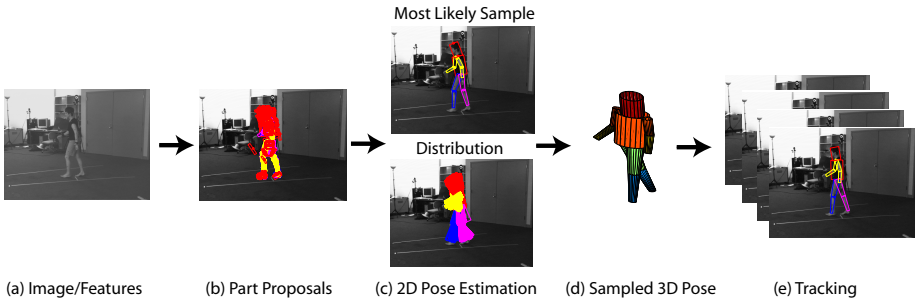


Fig. 1. Example of the hierarchical inference process. (a) monocular input image with bottom up limb proposals overlaid **(b)**; **(c)** distribution over 2D limb poses computed using non-parametric belief propagation; **(d)** sample of a 3D body pose generated from the 2D pose; **(e)** illustration of tracking.

To infer 2D body pose we adopt an iterative bottom-up process. Simple body part detectors provide noisy probabilistic proposals for the location and 2D pose (orientation and foreshortening) of visible limbs (Fig. 1 **(b)**). To estimate the pose of the body we exploit the idea of a 2D "loose-limbed" body model [20] which has been previously used for 2D articulated pose estimation [21] and 3D pose estimation and body tracking [20]. In particular, we adopt the view-based approach of [21]. We use a variant of non-parametric belief propagation (NBP) [8, 25] to infer probability distributions representing the belief in the 2D pose of each limb (Fig. 1 **(c)**). The inference algorithm also introduces hidden binary occlusion variables and marginalizes over them to account for occlusion relationships between body parts. The conditional distributions linking 2D body parts are learned from examples.

This process (limb proposals, NBP) provides reasonable guesses for 2D body pose from which to estimate 3D pose. Agarwal and Triggs [1, 2] learned a probabilistic mapping from 2D silhouettes to 3D pose using a Mixture of Experts (MoE) model. We generalize their approach to learn a mapping from 2D poses (including joint angles and foreshortening information) to 3D poses. Sampling from this model provides predicted 3D poses (Fig. 1 **(d)**), that are appropriate as proposals for a Bayesian temporal inference process (Fig. 1 **(e)**). Our multi-stage approach overcomes many of the problems inherent in inferring 3D pose directly from image features. We quantitatively evaluate the 3D proposals using ground truth 2D poses. We also test the method on the monocular sequence in Fig. 1.

2 Previous Work

There are now numerous methods for detecting the 2D pose of people in static images (with [5, 21] and without [7, 12–16] background subtraction). For example dynamic programming (DP) or other search methods can be used to compute possible 2D poses [5, 13–15]. While efficient DP methods exist [5], they require a discretization of the state space of 2D limb poses and simple forms for the conditional distributions relating connected limbs. They also require a tree structure, which does not allow long-range interactions between parts that are required for occlusion reasoning.

Alternatively, we adopt a graphical model representation of the body [21] that, in addition to kinematic constraints, also encodes the possible occlusion relationships between limbs (this leads to loops in the graph representation of the body). Pose estimation is formulated as inference in this loopy graphical model and is solved using a variant of Non-parametric Belief Propagation (NBP) [8, 25]. This leads to a number of advantages over DP methods. For example, limb positions and orientations need not be discretized as in [5]. Unlike previous methods [5, 21] we infer 2D pose as an intermediate step to inferring the full 3D articulated body pose.

Lee and Cohen [11] also use a bottom-up proposal process and infer 3D pose parameters using a data-driven MCMC procedure. Our approach differs in that we break the problem into simpler pieces: generate 2D proposals, inference of 2D pose, and prediction from 2D to 3D.

This final stage has received a good deal of attention with a variety of geometric [13, 26] and machine learning methods [1, 2, 17, 19, 22] being employed. These previous approaches have focused on directly inferring 3D pose from 2D silhouettes which may be difficult to obtain in general. Additionally silhouettes contain less information than our 2D models which represent all the limbs, the joint angles, and foreshortening. This helps reduce the ambiguities found in matching silhouettes to 3D models [23] but does not remove ambiguities altogether. Consequently we learn a conditional distribution using a MoE model similar to that of Agarwal and Triggs [1, 2]. Our work is similar in spirit to [6] in which 3D poses are inferred from 2D tracking results, but our approach can infer 3D pose from a single image and does not require manual initialization.

3 Modeling a Person

We model a 3D human body using a set of P (here $P = 10$) tapered cylinders corresponding to body parts and connected by revolute joints (see Fig. 3 (**a**)). Each part has an associated set of fixed parameters that are assumed to be known (e.g. length and cross-sectional radius at the two joints). We represent the overall pose of the body $Y_t = [\Xi, \Gamma, \theta]^T$ at time t using a set of joint angles θ, a global position Ξ, and global orientation Γ in 3D. Joint angles are represented with respect to the kinematic chain along which they are defined using unit quaternions. For our body model, this results in $Y_t \in R^{47}$, or $Y_t \in R^{55}$ depending on whether one chooses to model the clavicle joints.

In 2D the limbs in the image plane are modeled by trapezoids, and the overall body pose is defined using a redundant representation $X = \{X_1, X_2, ..., X_P\}$ in terms of 2D position, rotation, scale and foreshortening of parts, $X_i \in R^5$. This redundant representation stems from the inference algorithm that we will employ to infer the pose of the body in 2D. Notice we drop the temporal sub-script t for convenience.

4 Finding a Person in 2D

4.1 Limb Proposals

At the lowest level of our hierarchical approach are body part proposals. We need plausible poses/states for some or all the parts of the body to localize the search. There

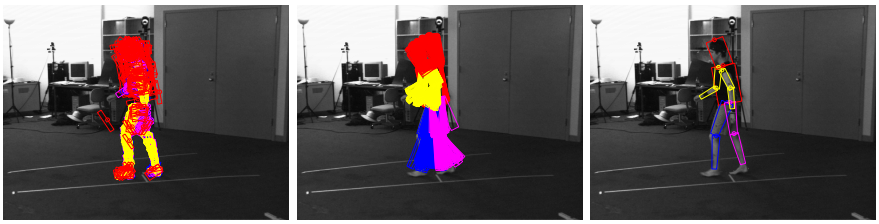


Fig. 2. Proposals and NBP. Example of the belief propagation process. Left: bottom-up proposals for the limbs. Center: 100 samples from the belief at each node/limb after 5 iterations of NBP (NBP was run with 100 particles, producing messages represented by 800-component kernel densities). Right: most likely sample drawn form the belief at each node.

exist a number of approaches for detecting body parts in an image. Among them are approaches for face detection, skin-color-based limb segmentation [11], and color-based segmentation exploiting the homogeneity and the relative spatial extent of body parts [11, 13, 16]. In this paper we took a simpler approach, and constructed our set of proposals by simply discretizing the state space and evaluating the likelihood function (below) at these discrete locations, choosing the 100 most likely states as a particle based proposal distribution for belief propagation (BP). It is important to note that not all parts need to be detected. An example of the proposals for various parts of the body are shown in Fig. 1 **(b)** and 2.

4.2 Likelihoods

The likelihood model for an individual limb is built to account for possible occlusions between body parts for a given view-based 2D model. To simplify the occlusion reasoning as in [21], we assume that for a given view there is a fixed and known depth ordering of parts. Assuming pixel independence, we can then write the local image likelihood $\phi(I|X_i)$, for part i as a product of individual pixel probabilities defined over disjoint image regions. For a more detailed description of the occlusion-sensitive likelihoods, and how one can approximate the global likelihood $\phi(I|X)$ with a product of local terms $\phi(I|X_i)$, we refer the reader to [21, 24]. In defining $\phi(I|X_i)$ we use silhouette and color features and combine them using an independence assumption.

4.3 2D Loose-Limbed Body Model

Following the framework of [20, 21] we implement the search for the 2D body using a spatial undirected graphical model, where each node i in a graph represents a body part (limb), and links between nodes represent the kinematic and occlusion constraints encoded statistically using conditional distributions. Each body part has an associated state vector $X_i \in R^5$ that encodes 2D position, rotation, scale, and foreshortening. The joint probability for this spatial graphical model with P body parts, can be written as $p(X_1, X_2, ..., X_P|I, V) \propto \prod_{ij} \psi_{ij}^K(X_i, X_j|V) \prod_{ij} \psi_{ij}^O(X_i, X_j|V) \prod_i \phi(I|X_i)$, where X_i represents the state of the limb i; $V \in \{1..8\}$ the discrete view; $\psi_{ij}^K(X_i, X_j|V)$ and $\psi_{ij}^O(X_i, X_j|V)$ are the kinematic and occlusion constraints between the connected

or potentially occluding nodes i and j for view V and $\phi(I|X_i)$ is the local image likelihood defined above. This model has a number of advantages [21] and has been shown to produce favorable results for the 3D body estimation in a multi-view setting [20]. The graphical model structure corresponding to our model can be seen in Fig. 3 (**b**).

Inferring the state of the 2D body in our graphical model representation corresponds to estimating the belief (marginal) at each node in a graph. We use a form of continuous non-parametric belief propagation [8], Particle Message Passing (PAMPAS), to deal with this task. The approach is a generalization of particle filtering which allows inference over arbitrary graphs rather then a simple chain. In this generalization the message used in standard belief propagation is approximated using a kernel density (formed by propagating particles through a conditional density represented by a mixture model [20, 21]). For the details on how the message updates can be carried out using the stratified sampling from the products of messages and proposal distribution see [20].

5 Proposing 3D body model from 2D

In order to produce estimates for the body in 3D from the 2D body poses, we need to model the conditional distribution $p(Y|X)$ of the 3D body state Y given 2D body state X. Intuitively this conditional mapping should be related to the inverse of the camera projection matrix and, as with many inverse problems, is highly ambiguous.

To model this non-linear relationship we use a Mixtures of Experts (MoE) model to represent the conditionals [1, 2, 22]. The parameters of the MoE model are learned by maximizing the log-likelihood of the training data set $D = \{X^1, ..., X^N, Y^1, ..., Y^N\}$ consisting of N input-output pairs (X^i, Y^i). We use an iterative Bayesian EM algorithm, based on type-II maximum likelihood, to learn parameters of the MoE. Our model for the conditional can be written as:

$$p(Y|X) \propto \sum_{k=1}^{M} p_{e,k}(Y|X, \Theta_{e,k}) p_{g,k}(k|X, \Theta_{g,k}) \tag{1}$$

where $p_{e,k}$ is the probability of choosing pose Y given the input X according to the k-th expert, and $p_{g,k}$ is the probability of that input being assigned to the k-th expert using an input sensitive gating network; in both cases Θ represents the parameters of the mixture and gate distributions.

For simplicity and to reduce complexity of the experts we choose linear regression with constant offset $Y = AX + C$ as our expert model, which allows us to solve for the parameters $\Theta_{e,k} = \{A_k, C_k, \Lambda_k\}$ analytically using the weighted linear regression, where $p_{e,k}(Y|X, \Theta_{e,k}) = \frac{1}{\sqrt{(2\pi)^n |\Lambda_k|}} \exp^{-\frac{1}{2}\Delta_k^T \Lambda_k^{-1} \Delta_k}$, and $\Delta_k = Y - A_k X - C_k$.

Pose estimation is a high dimensional and ill-conditioned problem, so simple least squares estimation of the linear regression matrix parameters typically produces severe over-fitting and poor generalization. To reduce this, we add smoothness constraints on the learned mapping. We use a damped regularization term $R(A) = \lambda ||A||^2$ that penalizes large values in the coefficient matrix A, where λ is a regularization parameter. Larger values of λ will result in overdamping, where the solution will be underestimated, small values of λ will result in overfitting and possibly ill-conditioning. Since

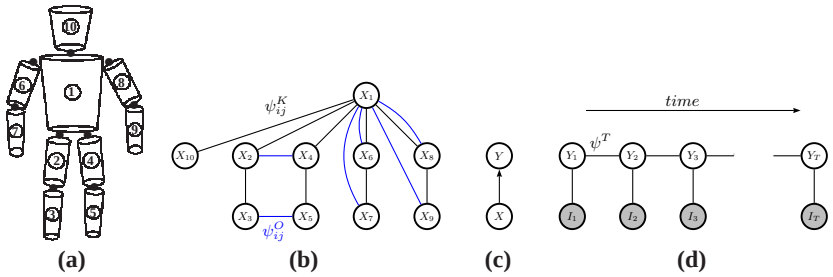


Fig. 3. Hierarchical Inference. Graphical model representation of the hierarchical inference process; **(a)** illustrates the 3D body model; **(b)** the corresponding 2D body model used for inference of the 2D pose at every frame, with kinematic constraints marked in black, and occlusion constraints in blue, and **(d)** the Hidden Markov Model (HMM) used for inferring and tracking the state of the 3D body, Y_t, over time $t \in \{1..T\}$, using the hierarchical inference proposed, in which proposals for each node, Y, are constructed from 2D body pose X using the model in **(c)**.

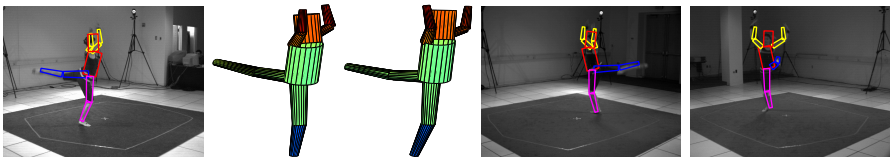
the solution of the ridge regressors is not symmetric under the scaling of the inputs, we normalize the inputs $\{X^1, X^2, ..., X^N\}$ by the standard deviation in each dimension respectively before solving [1]. We omit the details of weighted ridge regression due to space limitations, and refer readers to [2, 22].

Maximization for the gate parameters can be done analytically as well. Given the gate model, $p_{g,k}(k|X, \Theta_{g,k}) = \frac{1}{\sqrt{(2\pi)^n |\Sigma_k|}} \exp^{-\frac{1}{2}(X-\mu_k)^T \Sigma_k^{-1}(X-\mu_k)}$ maximization of the gate parameters $\Theta_{g,k} = (\Sigma_k, \mu_k)$ becomes similar to the mixture of Gaussians estimation, where $\mu_k = \sum_{n=1}^N z_k^n X^n / \sum_{n=1}^N z_k^n$, $\Sigma_k = \frac{1}{\sum_{n=1}^N z_k^n} \sum_{n=1}^N z_k^n (X^n - \mu_k)(X^n - \mu_k)^T$, and z_k^n is the the estimated ownership weight of the example n by the expert k estimated by expectation $z_k^n = \frac{p_{e,k}(Y^n|X^n, \Theta_{e,k}) p_{g,k}(k|X^n, \Theta_{g,k})}{\sum_{j=1}^M p_{e,j}(Y^n|X^n, \Theta_{e,j}) p_{g,j}(j|X^n, \Theta_{g,j})}$.

The above outlines the full EM procedure for the MoE model. We learn MoE models for two classes of actions: walking and dancing. Examples of the ground truth 2D query pose with corresponding expected 3D body pose can be seen in Fig. 4 **(a)** and **(b)** respectively. Similar to [1, 2] we initialize the EM learning by clustering the output 3D poses using the K-means procedure.

Implementation Details. Instead of learning the full conditional model $p(Y|X)$, we learn two independent models $p(\Gamma|X)$ and $p(\theta|X)$ one for the pose of the 3D body $p(\theta|X)$ given the 2D body pose X, and one for the global orientation of the body $p(\Gamma|X)$. The reasoning for this is two fold. First, this partitions the learned mapping into a fully camera-independent model for the pose $p(\theta|X)$, and the more specific camera-dependent model for the orientation of the body in the world $p(\Gamma|X)$. Second, we found that the optimal damping coefficient is significantly different for the two

[1] To avoid problems with 2D and 3D angles that wrap around at 2π, we actually regress the $(cos(\theta), sin(\theta))$ representation for 2D angles and unit quaternion $q = (x, y, z, w)$ representation for 3D angles. After the 3D pose is reconstructed we normalize the not-necessarily normalized quaternions to valid 3D rotations. Since quaternions also suffer from the double cover problem, where two unit quaternions correspond to every rotation, care must be taken to ensure that consistent parameterization is used.



(a) **(b)** **(c)** **(d)** **(e)**

Fig. 4. Proposed 3D pose. **(a)** Query 2D body pose; **(b)** expected 3D pose produced by the learned Mixture of Experts (MoE) model. **(c)** Ground-truth 3D body pose; **(d)** and **(e)** illustrate the projection of the expected 3D pose shown in **(b)** onto two alternative image views.

models that imposing a single joint conditional model (and hence a single coefficient) would result in somewhat larger reconstruction error. Estimation of the depth $p(\Xi|X)$ is done analytically by considering the estimated overall scale of the 2D body.

6 Tracking in 3D

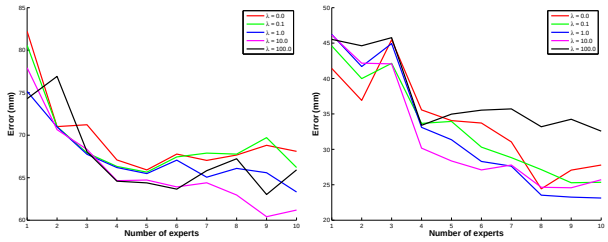
Once the distribution for the 3D body pose at every frame is inferred using the conditional MoE model described, we can incorporate temporal constraints to regularize the individual 3D pose estimates by tracking. We exploit the relatively standard [10] Hidden Markov Model (HMM) shown in Fig. 3 **(d)**. To infer the state of Y_t at every frame t given the temporal constraints $\psi^T(Y_t|Y_{t+1}) = \psi^T(Y_{t+1}|Y_t) \sim N(0, \Sigma_T)$ with learned covariance matrix Σ_T, we use the same inference framework of Non-parametric BP introduced in Section 4.3. Unlike many competing approaches, we allow the model to optimize the pose estimates not only forward but also backward in time in a batch.

The likelihood, $\phi(I_t|Y_t)$, of observing the 3D pose Y_t at time t given image evidence I_t is defined in terms of Chamfer distance of the projected pose Y_t to the silhouettes and edges obtained from I_t using standard techniques. Further details are omitted, and the reader is referred to [4] and [23] for similar likelihood model formulations.

7 Experiments

Datasets. For all experiments presented in this paper we used two datasets that exhibit two different types of actions: **walking** and **dancing**. Both datasets contain a number of motion capture examples used for training, and a single synchronized motion capture example with multi-view video used for testing. Video was captured using 4 stationary grayscale cameras at 60 Hz, and 3D pose was captured using a Vicon system at 120 Hz. The motion capture (mocap) was aligned to video and sub-sampled to 60 Hz, to produce synchronous video/mocap streams. All cameras were calibrated using standard calibration procedures. **Walking** dataset [20] contains 4587 training and 1398 testing poses/frames; **dancing**: 4151 training and 2074 testing poses/frames.

Quantitative evaluation of 2D to 3D pose mapping. Learning the mapping from 2D kinematic pose to 3D kinematic pose is one of the key contributions of this paper. We learned two action-specific MoE models $p(Y|X)$. For each of the action types we first looked at how sensitive our learned mapping is to the parameters of the model (i.e. the



(a) **(b)**

Fig. 5. Quantitative evaluation of action-specific **dancing** conditional model $p(Y|X) = p(\Xi|X)p(\Gamma|X)p(\theta|X)$, computed by comparing the expectation of the **(a)** 3D pose $E[p(\theta|X)]$, and of the **(b)** global orientation $E[p(\Gamma|X)]$ to ground truth data. Error is averaged over 4 trained MoE models learned with parameters specified. In both cases, **(a)** and **(b)**, it is clear that there is benefit in using large number of mixture components (> 5), and a moderate value for λ.

number of mixture components, and the regularization term λ). The results for **dancing** can be seen in Fig. 5. To quantitatively evaluate the performance we use the measure of [20] computed by choosing 15 virtual markers corresponding to joints and "ends" of limbs, and computing an expected absolute distance in (mm) over all the markers. Once the optimal set of parameters was chosen, the resulting MoE models were applied to the test data, and the error for the reconstructed 3D poses[2] analyzed (see Fig. 6).

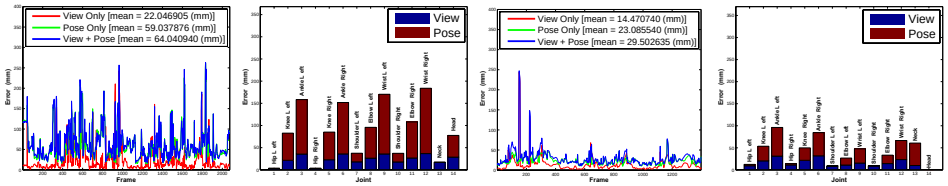
The key observation is that **walking**, being considerably simpler of the two action types, can be recovered significantly better (with 50% less error), than the more complex **dancing**. The peaks in the error in both cases often correspond to singular or close to singular cases where foreshortening in the pose of 2D limbs for example is severe.

Hierarchical inference from monocular image sequence. We also tested the full hierarchical inference on the first 50 frames from the **walking** test sequence. The 3D proposals obtained using the hierarchical inference process (Fig. 7) are accurate, and sufficient to allow reliable Bayesian temporal inference (Fig. 8).

8 Summary and Conclusions

The automatic estimation of human pose and motion in monocular image data remains a challenging problem. This is particularly so in the unconstrained environment where good background subtraction is unavailable. Here we have proposed a system to address this problem that uses a hierarchal Bayesian inference framework to go from crude body part detections to a distribution over 3D body pose. We make modest assumptions about the availability of noisy body part detectors and a reasonable image likelihood model. We use belief propagation to infer 2D limb poses that are consistent with the human body model. Our approach extends recent work on inferring 3D body models from 2D silhouettes by using the inferred 2D articulated model instead. This provides a richer representation which reduces ambiguities in the 2D to 3D mapping. We also show that the 3D pose proposals can be used in a tracking framework, that can further regularize the 3D pose estimates.

[2] Supplementary videos are available from `http://www.cs.brown.edu/people/ls/`



(a) **(b)** **(c)** **(d)**

Fig. 6. Quantitative evaluation of action-specific conditional model $p(Y|X) = p(\Xi|X)p(\Gamma|X)p(\theta|X)$, computed by comparing the expectation to ground truth data for two classes of motion. Per frame error for the reconstructed 3D poseθ, global orientation Γ, and the full 3D state of the body Y are shown for **(a) dancing** and **(c) walking**; the average per joint error as compared to the ground truth is shown in **(b)** and **(d)** respectively.

References

1. A. Agarwal and B. Triggs. Learning to track 3D human motion from silhouettes. *ICML*, pp. 9–16, 2004.
2. A. Agarwal and B. Triggs. 3D human pose from silhouettes by relevance vector regression. *CVPR*, vol. 2, pp. 882–888, 2004.
3. A. Balan, L. Sigal and M. Black. A quantitative evaluation of video-based 3D person tracking. *VS-PETS*, pp. 349–356, 2005.
4. J. Deutscher and I. Reid. Articulated body motion capture by stochastic search. *IJCV*, 61(2):185–205, 2004.
5. P. Felzenszwalb and D. Huttenlocher. Pictorial structures for object recognition. *IJCV*, 61(1):55-79, Jan. 2005.
6. N. R. Howe, M. E. Leventon and W. T. Freeman. Bayesian reconstruction of (3D) human motion from single-camera video. *NIPS*, pp. 820–826, 1999.
7. G. Hua, M.-H. Yang and Y. Wu. Learning to estimate human pose with data driven belief propagation. *CVPR*, vol. 2, pp. 747–754, 2005.
8. M. Isard. Pampas: Real-valued graphical models for computer vision. *CVPR*, vol. 1, pp. 613–620, 2003.
9. S. Ju, M. Black and Y. Yacoob. Cardboard people: A parametrized model of articulated motion. *Int. Conf. on Automatic Face and Gesture Recognition* , pp. 38–44, 1996.
10. X. Lan and D. Huttenlocher. A unified spatio-temporal articulated model for tracking. *CVPR*, vol. 1, pp. 722–729, 2004.
11. M. Lee and I. Cohen. Proposal maps driven MCMC for estimating human body pose in static images. *CVPR*, vol. 2, pp. 334–341, 2004.
12. G. Mori. Guiding model search using segmentation. *ICCV*, pp. 1417–1423, 2005.
13. G. Mori, X. Ren, A. Efros and J. Malik. Recovering human body configurations: Combining segmentation and recognition. *CVPR* vol. 2, pp. 326–333, 2004.
14. D. Ramanan, D. Forsyth and A. Zisserman. Strike a pose: Tracking people by finding stylized poses. *CVPR*, vol. 1, pp. 271–278, 2005.
15. D. Ramanan and D. Forsyth. Finding and tracking people from the bottom up. *CVPR*, vol. 2, pp. 467–474, 2003.
16. T. Roberts, S. McKenna and I. Ricketts. Human pose estimation using learnt probabilistic region similarities and partial configurations. *ECCV*, vol. 4, pp. 291–303, 2004.
17. R. Rosales and S. Sclaroff. Inferring body pose without tracking body parts. *CVPR*, vol. 2, pp. 721–727, 2000.
18. H. Sidenbladh, M. Black and D. Fleet. Stochastic tracking of 3D human figures using 2D image motion. *ECCV*, vol. 2, pp. 702–718, 2000.

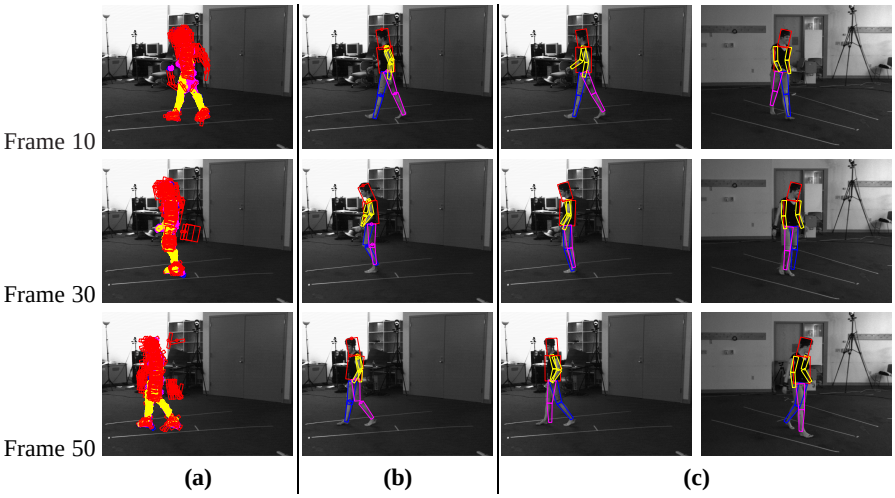


Fig. 7. Hierarchical 3D Pose Estimation. **(a)** bottom-up proposals for the limbs, **(b)** most likely sample from the marginals for each limb after 2D pose estimated by NBP, and **(c)** most likely 3D pose obtained by propagating 2D poses through a conditional $p(Y|X)$ model.

19. G. Shakhnarovich, P. Viola and T. Darrell. Fast pose estimation with parameter-sensitive hashing. *ICCV*, vol.2, pp. 750–759, 2003.
20. L. Sigal, S. Bhatia, S. Roth, M. Black and M. Isard. Tracking loose-limbed people. *CVPR*, vol. 1, pp. 421–428, 2004.
21. L. Sigal and M. Black. Measure Locally, Reason Globally: Occlusion-sensitive articulated pose estimation. *CVPR*, 2006.
22. C. Sminchisescu, A. Kanaujia, Z. Li and D. Metaxas. Discriminative density propagation for 3D human motion estimation. *CVPR*, vol. 1, pp. 390–397, 2005.
23. C. Sminchisescu and B. Triggs. Estimating articulated human motion with covariance scaled sampling. *IJRR*, 22(6), pp. 371–391, 2003.
24. E. Sudderth, M. Mandel, W. Freeman and A. Willsky. Distributed occlusion reasoning for tracking with nonparametric belief propagation. *NIPS*, pp. 1369–1376, 2004.
25. E. Sudderth, A. Ihler, W. Freeman and A. Willsky. Nonparametric belief propagation. *CVPR*, vol. 1, pp. 605–612, 2003.
26. C. J. Taylor. Reconstruction of articulated objects from point correspondences in a single image. *CVIU*, 80(3):349–363, 2000.

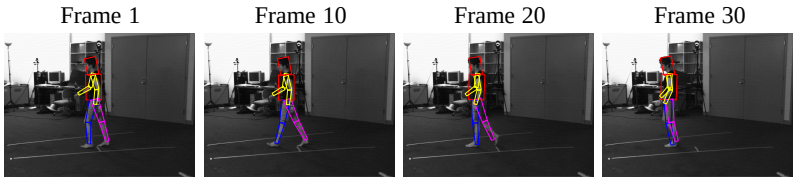


Fig. 8. Tracking in 3D. Tracking based on the 3D proposals (Fig. 7) at 10 frame increments. The 3D poses are projected into the image for clarity. The mean tracking error of 66 (*mm*), computed over first 50 frames of the test sequence, is 77% lower then the error reported for the same dataset using single-view Annealed Particle Filter (APF) with manual initialization in [3]. The best reported result in the literature on this data of 41 (*mm*) was obtained using 4-view APF [3].