

Monocular Tracking with a Mixture of View-Dependent Learned Models

Tobias Jaeggli¹, Esther Koller-Meier¹, and Luc Van Gool^{1,2}

¹ ETH Zurich, D-ITET/BIWI, CH-8092 Zurich

² Katholieke Universiteit Leuven, ESAT/VISICS, B-3001 Leuven
jaeggli@vision.ee.ethz.ch

Abstract. This paper considers the problem of monocular human body tracking using learned models. We propose to learn the joint probability distribution of appearance and body pose using a mixture of view-dependent models. In such a way the multimodal and nonlinear relationships can be captured reliably. We formulate inference algorithms that are based on generative models while exploiting the advantages of a learned model when compared to the traditionally used geometric body models. Given static images or sequences, body poses and bounding box locations are inferred using silhouette based image descriptors. Prior information about likely body poses and a motion model are taken into account. We consider analytical computations and Monte-Carlo techniques, as well as a combination of both. In a Rao-Blackwellised particle filter, the tracking problem is partitioned into a part that is solved analytically, and a part that is solved with particle filtering. Tracking results are reported for human locomotion.

1 Introduction

Bayesian approaches have been successfully applied to human body tracking. Typically, these approaches are generative and need a mechanism to predict a subject’s appearance given hypotheses for the parameters that are to be estimated.

Previous tracking algorithms often work with hand-crafted geometric body models that are rendered and compared to input images in order to verify body pose hypotheses (e.g. [1,2,3,4,5]). These body models have many parameters such as limb lengths and widths that all have to be known or estimated, typically in an initialisation procedure or even on-the-fly.

As opposed to geometrical models, probabilistic Machine Learning methods naturally offer the possibility to learn the dependencies of body pose and its appearance while generalising over irrelevant variation of appearance and inter-person variance.

Core of the proposed approach is a model of the statistical dependencies between body poses and their appearance, which is learned from training data. We show how the learning problem itself can be alleviated by uncoupling global orientation and local body pose, and learn the joint distribution over pose and appearance as a mixture of view-dependent models. Within a view-dependent model, the distribution is captured by the means of Gaussian Mixture Models (GMMs).

The Recursive Bayesian Filter serves as an overall framework for inference. Appearance is encoded using image descriptors that are computed from the silhouette of

background segmented images. Silhouettes provide rich information about body pose, but leave certain aspects unobservable, *i.e.* are subject to ambiguities. This enforces the use of prior information on one hand, and will lead to multimodal posteriors on the other hand.

1.1 Related Work

There is a wide variety of literature about probabilistic body tracking. Most methods use geometric body models (*e.g.* [1,2,3,4]), sometimes in conjunction with learning techniques for dimensionality reduction or to estimate parameters of the prior or observation model. Example based approaches (*e.g.* [6]) are often based on nearest neighbour search and provide mechanisms for efficient lookup in large databases. Several authors have applied parametric machine learning methods to body pose estimation [7,8,9,10,11] and aim at learning the relationship between image observations and body pose, which is challenging because the mapping is nonlinear and multivalued. The further discussion will concentrate on these works.

In [7] the authors assume a functional relationship between silhouette descriptors and pose and propose relevance vector regression for learning. Grauman *et al.* [10] learn a density over multiple silhouettes and corresponding structure using a mixture of PPCA. Given a (static) set of silhouettes, the MAP estimate is obtained. Recently, inference algorithms that explicitly deal with multimodal posterior distributions have been applied to the body tracking problem. In the specialized mappings architecture [12], multiple functional mappings from visual features to articulated pose are learned. Inference yields a set of hypotheses, a problem specific method is then used to compute the likelihood of the different hypotheses. Most related to our work are [8] and [11] that both learn the conditional pdf of pose and appearance with a mixture of regressors (experts).

In [11] the temporal dependencies and image-pose dependencies are learned in a single discriminative model. The distributions are propagated analytically. In [8], a pdf over possible poses is inferred given an input silhouette. The analytical inference procedure does not include any temporal aspects. For tracking, a particle filter in high dimensions is used, where the inferred pdf is treated as the observation likelihood. The algorithms proposed in this paper follow well known generative formulations, however we propose solutions that are based both on sampling techniques and analytical inference where applicable, thereby avoiding the need to sample in high dimensions.

To summarise, this paper mainly contributes by explicitly addressing the issue of learning appearance from all view directions while allowing for multimodalities of the distribution. Analytic solutions to the generative formulation of the tracking problem are proposed, and a Rao-Blackwellised particle filter that combines the advantages of sample-based and analytic inference. Furthermore, since we learn the joint pdf of pose and appearance, rather than the conditional, we can use this model to estimate the 2d image position of the bounding box along with the body pose, an issue not addressed in [8,11].

The paper is organised as follows. Section 2 introduces the mixture of view-dependent models, and 3 formulates tracking algorithms based on the learned models. In 4 we describe our implementation and show experimental results and conclude in sect. 5.

2 Mixture of view-dependent models

We want to learn the dependencies of body pose and its appearance in images. The state space for the body pose is given by the variables α and \mathbf{x} , the global orientation of the body relative to the camera and its local pose, *i.e.* the configuration of its limbs. Under the assumption that the camera is in an approximatively horizontal position, at face or shoulder level, the global orientation can be described with a single parameter that determines the position on a circle around the object from which the latter is observed. We therefore face the problem of learning the joint pdf $p(\alpha, \mathbf{x}, \mathbf{y})$, where \mathbf{y} is an observation, *i.e.* a descriptor that is computed the input image. In order to simplify the learning problem, we rewrite this pdf as a mixture of C view-dependent models p_c that each cover a section of possible view directions/global orientations.

$$p(\alpha, \mathbf{x}, \mathbf{y}) = \frac{1}{C} \sum_{c=1}^C p_c(\alpha, \mathbf{x}, \mathbf{y}) \quad (1)$$

Within the view-dependent models, there is little variation of view direction, so the view angle can be assumed independent from local pose and observation, which enables us to rewrite equation (1) as

$$p(\alpha, \mathbf{x}, \mathbf{y}) = \frac{1}{C} \sum_{c=1}^C p_c(\alpha) p_c(\mathbf{x}, \mathbf{y}), \quad (2)$$

where $p_c(\alpha)$ is a one-dimensional Gaussian $\mathcal{N}(\alpha; \alpha_c, \sigma)$ ¹ and $p_c(\mathbf{x}, \mathbf{y})$ is the joint pdf of pose and appearance for a certain view direction; this pdf will be learned from training data. Within a view-dependent model the view angles are normally distributed around the mean α_c , with α_c 's uniformly spaced over the interval $[0, 2\pi[$, and variances chosen such that adjacent models overlap, and the whole domain of α is uniformly covered.

The view-dependent models $p_c(\mathbf{x}, \mathbf{y})$ themselves are approximated by a mixture of Gaussians (GMM), estimated using *e.g.* an EM algorithm. The joint distribution over orientation, pose and appearance is thus a mixture of mixtures of Gaussians.

$$p(\alpha, \mathbf{x}, \mathbf{y}) = \frac{1}{C} \sum_{c=1}^C \left[p_c(\alpha) \sum_{s=1}^S w_{c,s} \mathcal{N}(\mu_{c,s}, \Sigma_{c,s}) \right] \quad (3)$$

Here, S is the number of Gaussian components in each p_c , and $w_{c,s}$ are the weights estimated by the EM algorithm in the learning phase ($\sum_{s=1}^S w_{c,s} = 1$). $\mu_{c,s}$ and $\Sigma_{c,s}$ are the parameters of the Gaussian components.

Note that even though the omnidirectional model $p(\alpha, \mathbf{x}, \mathbf{y})$ consists of a discrete number of almost unidirectional models, we have defined a smooth and continuous overall model that covers the entire state space.

¹ We use the notation $\mathcal{N}(x; \mu, \sigma)$ for Gaussian distributions, where the first argument is omitted if clear from the context.

3 Tracking with learned models

According to Bayes' rule, the tracking problem can be formulated as

$$p(\mathbf{X}_t | \mathbf{y}_{1:t}) \propto p(\mathbf{y}_t | \mathbf{X}_t) p(\mathbf{X}_t | \mathbf{y}_{1:t-1}), \quad (4)$$

where \mathbf{X}_t is the state variable we want to infer from aggregated observations $\mathbf{y}_{1:t}$ ($\mathbf{X}_t = [\mathbf{x}_t, \alpha_t]^T$ for the notation of Sect. 2). The image likelihood is obtained from our learned model of $p(\mathbf{X}, \mathbf{y})$ by

$$p(\mathbf{y}_t | \mathbf{X}_t) = p(\mathbf{X}_t, \mathbf{y}_t) / p(\mathbf{X}_t). \quad (5)$$

As we have not learned the temporal prior $p(\mathbf{X}_t | \mathbf{X}_{t-1})$ explicitly, we would like to include information about likely body poses as well as a motion model in its definition. We model the temporal behaviour as a Brownian motion around the expected new position multiplied by the time independent prior $p(\mathbf{X}_t)$. See Fig. 1a) for an illustration.

$$p(\mathbf{X}_t | \mathbf{X}_{t-1}) := k(\mathbf{X}_{t-1}) p(\mathbf{X}_t) \mathcal{N}(A\mathbf{X}_{t-1}, \Sigma_T) \quad (6)$$

Here, A specifies the linear dependencies between subsequent states, and $k(\mathbf{X}_{t-1}) = \int_{\mathbf{X}_t} p(\mathbf{X}_t) \mathcal{N}(A\mathbf{X}_{t-1}, \Sigma_T)$ is a normalisation factor. Using this definition, we obtain

$$\begin{aligned} p(\mathbf{X}_t | \mathbf{y}_{1:t-1}) &= \int_{\mathbf{x}_{t-1}} p(\mathbf{X}_t | \mathbf{X}_{t-1}) p(\mathbf{X}_{t-1} | \mathbf{y}_{1:t-1}) \\ &= \int_{\mathbf{x}_{t-1}} k(\mathbf{X}_{t-1}) p(\mathbf{X}_t) \mathcal{N}(A\mathbf{X}_{t-1}, \Sigma_T) p(\mathbf{X}_{t-1} | \mathbf{y}_{1:t-1}). \end{aligned} \quad (7)$$

The factor $k(\mathbf{X}_{t-1})$ depends on \mathbf{X}_{t-1} which makes analytic integration intractable; it can however be computed explicitly in a sampling based approach. We propose a slightly different definition that is suitable for both analytic and Monte-Carlo integration.

$$p(\mathbf{X}_t | \mathbf{y}_{1:t-1}) := K p(\mathbf{X}_t) \int_{\mathbf{x}_{t-1}} \mathcal{N}(A\mathbf{X}_{t-1}, \Sigma_T) p(\mathbf{X}_{t-1} | \mathbf{y}_{1:t-1}) \quad (8)$$

This formulation corresponds to first propagating the old posterior according to the motion model, and then eliminating unlikely body poses. Both (7) and (8) define a pdf over \mathbf{X}_t that takes into account temporal as well as static prior information.

By combining (4), (5) and prior (8), \mathbf{X}_t can now be inferred analytically for any given sequence.

$$\begin{aligned} p(\mathbf{X}_t | \mathbf{y}_{1:t}) &\propto p(\mathbf{y}_t | \mathbf{X}_t) p(\mathbf{X}_t) \int_{\mathbf{X}_{t-1}} \mathcal{N}(\mathbf{X}_t; A\mathbf{X}_{t-1}, \Sigma_T) p(\mathbf{X}_{t-1} | \mathbf{y}_{1:t-1}) \\ &= p(\mathbf{X}_t, \mathbf{y}_t) \int_{\mathbf{X}_{t-1}} \mathcal{N}(\mathbf{X}_t; A\mathbf{X}_{t-1}, \Sigma_T) p(\mathbf{X}_{t-1} | \mathbf{y}_{1:t-1}). \end{aligned} \quad (9)$$

In order to account for noisy observations, the term $p(\mathbf{X}_t, \mathbf{y}_t)$ is computed by multiplying the learned model with a Gaussian pdf around the actual observation for \mathbf{y}_t , denoted \mathbf{y}_{obs} , and then marginalising over \mathbf{y}_t . This is illustrated in Fig. 1b). Marginalisation of a

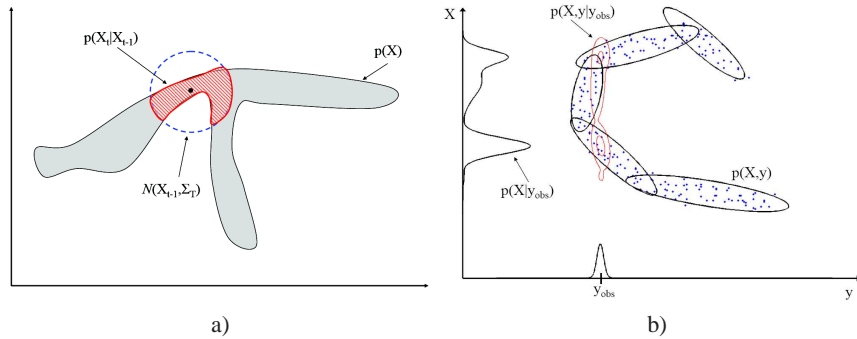


Fig. 1. a) The overall prior (red, hatched) is defined as the product of the motion model (blue, dashed, for this illustration a Gaussian pdf around the state of $t - 1$) and the static learned prior $p(\mathbf{X})$. b) Multiplication of the Gaussian distribution around the observation \mathbf{y}_{obs} and the prior $p(\mathbf{X}, \mathbf{y})$ yields $p(\mathbf{X}, \mathbf{y}|\mathbf{y}_{obs})$. By marginalisation, the pdf $p(\mathbf{X}|\mathbf{y}_{obs})$ over unobserved variables is then obtained. Note the multimodality of $p(\mathbf{X}|\mathbf{y}_{obs})$.

GMM is straightforward; the marginal mixture has the same number of Gaussian components as the original joint mixture with the same weights. The means and covariances of the marginal mixture are simply the means and covariances of the original mixture with all elements corresponding to the variable \mathbf{y} removed.

The integral in Eq. (9) can be calculated in closed form and will result in a Gaussian mixture, so the result of (9) is the product of two mixtures and thus a mixture itself. However, the number of mixture components will grow exponentially over time. Hence, at each timestep a mixture simplification step reduces the number of Gaussian components, by pruning components with very low weights and replacing clusters of components by their 'average' Gaussian.

3.1 Rao-Blackwellised particle filter

So far we assumed that an observation is available in the form of an image descriptor computed at a certain position in the image. This requires that the bounding box containing the person is either known on beforehand or estimated in some way. For the silhouette based image descriptor, one could imagine an ad-hoc algorithm for this 2d tracking problem. In general, however, we want to support multiple hypotheses for the 2d location variables \mathbf{l} , so they have to be included in our state space and inferred by the overall tracking algorithm. A particle based approach can easily be extended accordingly by adding these location-variables to the state space. In the case of an analytic approach however, there is no straightforward extension since the posterior over this extended state space is unlikely to have parametric form. We therefore propose to partition our state space into a part that is solved using a particle filter and a part that is solved analytically using our learned models. By the chain rule of probability, the posterior over \mathbf{x} , α , and the location variable \mathbf{l} can be written as

$$p(\mathbf{x}, \alpha, \mathbf{l}|\mathbf{y}) = p(\mathbf{x}|\alpha, \mathbf{l}, \mathbf{y})p(\alpha, \mathbf{l}|\mathbf{y}), \quad (10)$$

where the temporal aspects of the problem are omitted for notational simplicity. Given our learned model, $p(\mathbf{x}|\alpha, \mathbf{l}, \mathbf{y})$ can be inferred analytically and described parametrically, whereas for $p(\alpha, \mathbf{l}|\mathbf{y})$ no analytic solution is obvious. However, due to its low dimensionality, it can be handled by a particle filter. The Rao-Blackwellised Particle Filter (RBPF, [13]) offers a framework for inference, when a part of the state space can be marginalised analytically. Figure 2 a) shows the graphical structure of this setting as a Bayesian network.

In RBPF, each particle will consist of a sample for \mathbf{l}_t and α_t , a parametric pdf $p(\mathbf{x}_t|\mathbf{y}_{1:t}, \mathbf{l}_{1:t}^i, \alpha_{1:t}^i)$ and a weight w_t^i . The computation of $p(\mathbf{x}_t|\mathbf{y}_{1:t}, \mathbf{l}_{1:t}^i, \alpha_{1:t}^i)$ follows the general derivation for analytic density propagation (9), except that we only infer the variable \mathbf{x}_t , and that the expression is additionally conditioned on $\alpha_{1:t}^i$ and $\mathbf{l}_{1:t}^i$. We will denote as \mathbf{y}_t^i the image descriptor computed at sampled location \mathbf{l}_t^i .

$$\begin{aligned}
 p(\mathbf{x}_t|\mathbf{y}_{1:t}, \alpha_{1:t}^i) &= \frac{1}{L_i} p(\mathbf{y}_t^i|\mathbf{x}_t, \alpha_t^i) p(\mathbf{x}_t|\mathbf{y}_{1:t-1}, \alpha_{1:t-1}^i) \\
 &\propto \frac{K_i}{L_i} p(\mathbf{x}_t, \alpha_t^i, \mathbf{y}_t^i) \int_{\mathbf{x}_{t-1}} \mathcal{N}(\mathbf{x}_t; A\mathbf{x}_{t-1}, \Sigma_T) p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1}, \alpha_{1:t-1}^i)
 \end{aligned}
 \tag{11}$$

Here we used the independence of \mathbf{x}_t and α_t^i and the uniformity of $p(\alpha_t^i)$. K_i is the scaling factor from the prior (8), and the normalisation factor L_i is equal to the likelihood of the observation given the i th sample. Hence, if we choose the prior $p(\mathbf{l}_t, \alpha_t|\mathbf{l}_{t-1}, \alpha_{t-1})$ as a proposal function, the weights w_t^i are given by the normalisation factor L_i [13].

The RBPF harmonizes well with the mixture of view-dependent models; for the computation of $p(\mathbf{x}_t, \alpha_t^i, \mathbf{y}_t^i)$ in (11) we can exclude those mixture components that are not compatible with the hypothesis α_t^i , i.e that will have a very low weight in the posterior mixture. This decreases computation time per sample significantly.

4 Implementation and Experimentation

The previous sections are kept general with respect to the used image descriptors, the body pose parametrisation and the classes of motion for which the system is trained. In this section we present an implementation and experimental validation that serve as a proof of concept and aim at illustrating the potential of the overall approach. We chose human locomotion as a case in point, but expect that an extension to more general motions is feasible provided that such training data are available.

4.1 Image and Pose Descriptors

The chosen image descriptors are based on the silhouette of the tracked person. Using a stationary camera, the segmentation is obtained via background subtraction. To encode these segmented images using a descriptor of moderate size, we use signed distance functions, that assign to each pixel a signed value indicating the distance to the closest point on the silhouette [9]. These values are computed on a grid of equidistantly spaced sample points inside the bounding box of the segmented object. Several examples of such distance-transformed silhouettes are shown in Fig. 2 b).

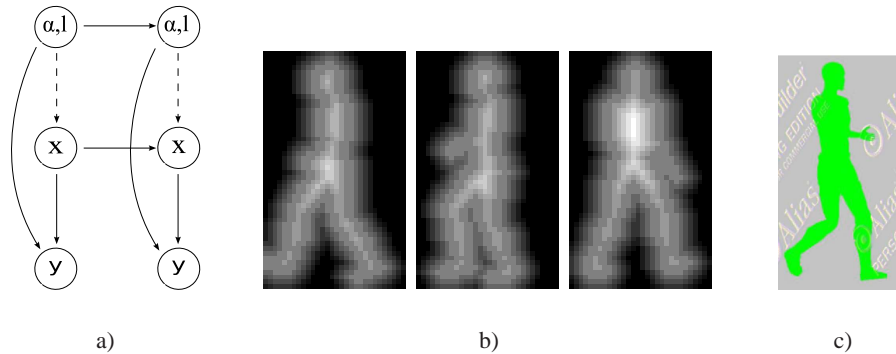


Fig. 2. a) Graphical structure obtained by partitioning the search space into two parts. This is the setting in which the Rao-Blackwellised particle filter operates. b) Signed distance functions as image descriptors. Positive values inside the silhouette, negative values outside, and 0 on the silhouette, rescaled here to the interval 0:255 for visualisation. c) Animated 3D body model (training data)

To compute this distance transform efficiently we use an algorithm similar to the chamfer image transform based on a hybrid distance measure that is an approximation to the real Euclidean distance. See [14] for an overview of algorithms.

Both image descriptors y and pose descriptors x are potentially high dimensional; this is a difficulty for the learning task. Furthermore we believe that the intrinsic dimensionality of the training data is much lower. A dimensionality reduction step is necessary. Here, we use PCA to bring down the dimensionality of both image and pose descriptor.

4.2 Experiments

To generate training data for this experiment, we rendered 11 MoCap² sequences from several subjects with different walking styles from 36 viewpoints using MotionBuilder PLE³, a package that is designed for realistic human animation. On the silhouettes of these renderings, the image descriptors were computed, followed by a PCA dimensionality reduction that retained the first 15 principal components. The body pose was represented using 3d joint locations for a number of joints that constitute the overall body pose (foot, knee, hip, shoulder, elbow, hand and head). Only the first 15 principal components, capturing about 99 percent of the variation, were retained for the final pose representation. For each view dependent model, the joint distribution of appearance and pose descriptors was approximated by a GMM with 11 components using an EM algorithm.

Using a plain particle filter in combination with this pose representation would require a large number of particles. Here we report results that were obtained using the algorithm from sect. 3.1, where a part of the inference problem is solved analytically. Samples for l and α are generated from a temporal prior that assumes constant velocity

² data obtained from <http://mocap.cs.cmu.edu/>

³ www.alias.com

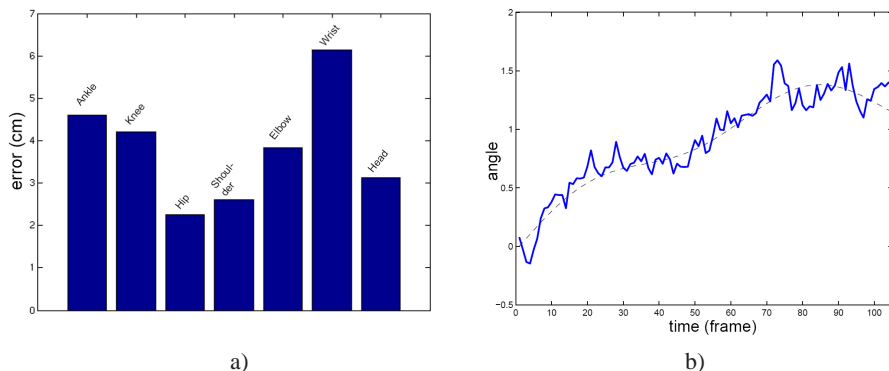


Fig. 3. a) Deviation from ground truth for a synthetic sequence. Euclidean distance (centimeters) between reconstructed joints and the ground truth, averaged over the sequence. The error is largest at the extremities (ankle, wrist). b) Estimated view directions α (radians) for a synthetic sequence with ground truth (dashed curve). The figure shows the angle encoded by the sample with the highest weight.

resp. Brownian motion. The temporal model for pose \mathbf{x} assumed Brownian motion in PCA-reduced pose space, with a covariance matrix learned from the training data. For the initialisation of the 2d location variables \mathbf{l} , an ad hoc proposal function at the center of gravity of the segmented image was used to sample from, \mathbf{x} and α were initialised using the analytical inference equation (9), by assuming a uniform temporal prior.

Some quantitative results are shown in Fig. 3 a), for a sequence that was synthetically generated the same way as the training data, but using MoCap from a subject not contained in the training set.

Figure 4 shows tracking through a real office sequence. The images were recorded with a DV camera at a frame rate of 25 fps and segmentation was obtained using background subtraction. The reconstructed poses and motion look natural. Occasionally (e.g. frame c), the reconstruction of the arms is imprecise, especially when they are occluded by the torso, i.e. not visible in the silhouette images. In such cases the pose prior alone is responsible for the estimation of the arm pose. Figure 3 b) shows the estimated view direction for a synthetic sequence with varying viewpoint. The reconstruction follows the overall rotation of the person. Largest deviation from ground truth is about 15 degrees (frame 73), the average error is 5 degrees. However there seems to be no systematic mis-estimation since the mean difference from ground truth is only 1 degree, the negative and positive deviations basically sum to zero. These results are very convincing, when considering that it is very difficult, even for humans, to perceive the relative orientation of a body from the silhouette alone.

5 Summary and Conclusion

We presented a system for monocular tracking of people. From a MoCap training database distributions over body pose and corresponding image appearance descriptors (silhouettes) are learned. Based on the learned model we were able to formulate

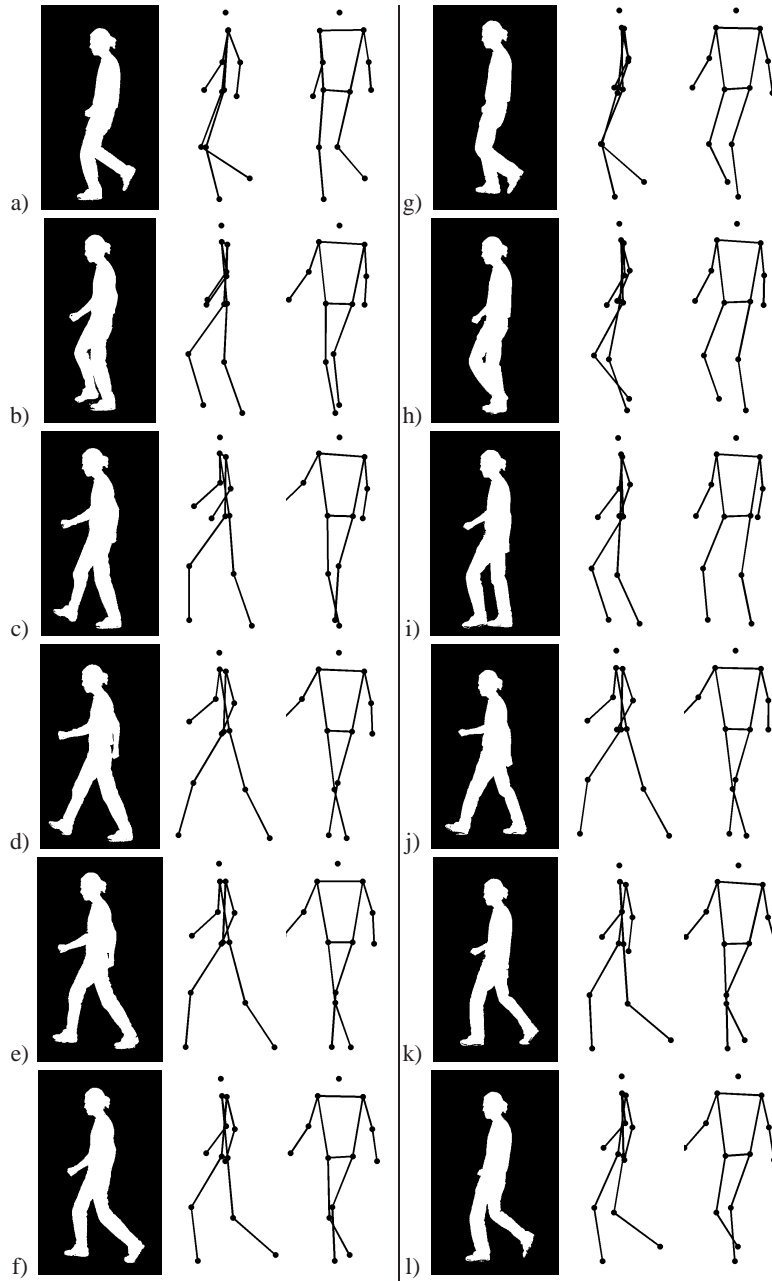


Fig. 4. Tracking through a real sequence. For each of the selected frames, the left column shows the tracked bounding box. The other columns show the estimated pose from side view resp. 45 degrees. To visualise a single pose per frame, we chose the mean of the component with the highest weight from the GMM that corresponds to the sample with the highest weight of the sample set. Note that between frame i) and j) the posterior mode that corresponds to stepping forward with the left leg suddenly becomes more likely, which can only be seen from the 45 degree view.

generative tracking algorithms that either work with analytical inference or particle sets or a combination (Rao-Blackwellised particle filter). Compared to approaches with geometrical models, we do not have to sample in high dimensions, and in contrast to purely discriminative learning based approaches, we can solve to 2d bounding box tracking along with the pose estimation. The algorithms were evaluated on synthetic and real sequences of walking people.

Future research directions will include the investigation of different image descriptors that do not require a foreground segmentation. Further experimental evaluation will focus on the multimodality of the posteriors that reflects the inherent ambiguities of the body tracking problem. We will also try to include a wider range of motions and actions into our models. Finally we aim at designing more elaborate temporal priors, possibly learned from the training data.

Acknowledgements

This work was supported by the SNF project PICSEL, the EU Integrated Project DIRAC, and the SNF NCCR IM2.

References

1. Deutscher, J., Blake, A., Reid, I.: Articulated body motion capture by annealed particle filtering. In Proc. IEEE CVPR (2000) [1](#), [2](#)
2. Sidenbladh, H., Black, M., Fleet, D.: Stochastic tracking of 3d human figures using 2d image motion. In ECCV (2000) 702–718 [1](#), [2](#)
3. Sigal, L., Bhatia, S., Roth, S., Black, M., Isard, M.: Tracking loose-limbed people. CVPR (2004) [1](#), [2](#)
4. Sminchisescu, C., Triggs, B.: Kinematic jump processes for monocular 3d human tracking. CVPR (2003) [1](#), [2](#)
5. Urtasun, R., Fua, P.: 3d human body tracking using deterministic temporal motion models. In ECCV (2004) [1](#)
6. Shakhnarovich, G., Viola, P., Darrel, T.: Fast pose estimation with parameter sensitive hashing. ICCV (2003) [2](#)
7. Agarwal, A., Triggs, B.: 3d human pose from silhouettes by relevance vector regression. In CVPR (2004) [2](#)
8. Agarwal, A., Triggs, B.: Monocular human motion capture with a mixture of regressors. IEEE Workshop on Vision for Human-Computer Interaction at CVPR (2005) [2](#)
9. Elgammal, A., Lee, C.S.: Inferring 3d body pose from silhouettes using activity manifold learning. CVPR (2004) [2](#), [6](#)
10. Grauman, K., Shakhnarovich, G., Darrel, T.: Inferring 3d structure with a statistical image-based shape model. In ICCV (2003) [2](#)
11. Sminchisescu, C., Kanaujia, A., Li, Z., Metaxas, D.: Discriminative density propagation for 3d human motion estimation. CVPR (2005) [2](#)
12. Rosales, R., Sclaroff, S.: Learning body pose via specialized maps. In Advances in Neural Information Processing Systems (2001) [2](#)
13. Murphy, K., Russel, S.: Rao-blackwellized particle filtering for dynamic bayesian networks. In A. Doucet, N. de Freitas and N. Gordon, editors, *Sequential Monte Carlo Methods in Practice*, pp 499-515, Springer (2001) [6](#)
14. Bailey, D.G.: An efficient euclidean distance transform. IWCIA (2004) [7](#)