Estimation of Human Figure Motion
Using Robust Tracking of Articulated Layers

Kooksang Moon and Vladimir Pavlović
Department of Computer Science
Rutgers University
Piscataway, NJ 08854

Abstract

We propose a probabilistic method for tracking articulated objects, such as the human figure, across multiple layers in monocular image sequence. In this method, each link of a probabilistic articulated object is assigned to one individual image layer. The layered representation allows us to robustly model the pose and occlusion of object parts during its motion. Appearance of links is described in terms of learned statistics of basic image features, such as color, and geometric models of robust spatial kernels. This results in a highly efficient computational method for inference of the object's pose. We apply this approach to tracking of the human figure in monocular video sequences. We show that the proposed method, coupled with a learned dynamic model, can lead to a robust articulated object tracker.

1. Introduction

Tracking and pose estimation of articulated objects, such as the human figure or hand, is a critical task in applications ranging from smart surveillance to advanced user interfaces. However, articulated objects often exhibit complex and dynamic behavior that makes motion tracking challenging. The task becomes increasingly difficult if only monocular image sequences are available because of frequent occlusions and shadowing among the moving object parts. A final impetus present cluttered scenes and, possibly, other moving objects.

Most articulated object tracking and pose estimation methods from monocular sequences have been studied in the context of the human figure. The approaches employed differ in focus between those based on a more complex 3D articulated human models [21, 6, 17, 16, 19, 20, 1] and simpler 2D figure-based methods [12, 3]. While computationally more efficient, the 2D model-based methods are often unable to effectively deal with the self-occlusions of the human figure, image and dynamic ambiguities and discontinuous motion.

In this paper we revisit the 2D-based articulated object tracking approach using a robust probabilistic, learning-based method. The crux of our method are: (a) a layered representation of the articulated object that approximates the true 3D link relationships and (b) a robust parametric statistical representation of the link appearances. The two representations are married together using a probabilistic graphical modeling formalism with a dynamic motion model representation. We assign each articulated object link to a unique 2D image layer. Within each layer the pose of the link is modeled using a robust kernel whose parameters encode the continuous 2D position, orientation, and scale of the link. We model the link's appearance using region-based feature statistics, such as the color histograms. The key advantages of this approach are in its computational efficiency and robustness to occlusions and ambiguities. Computational efficiency is induced by: (a) the region-based appearance representation, (b) continuous pose estimates that preclude the need for combinatorial search, (c) parametric density models, and (d) the layered representation that (seemingly) decouples the estimations of pose and appearance. Kernel-based representation of links introduces additional robustness to the approach.

This paper is organized as follows. Section 2 describes the related work in human motion tracking area. Section 3 introduce the new tracking framework and its probabilistic formulation, while in Section 4 we propose an efficient algorithm for pose estimation and tracking. Sections 5 and 6 presents our experimental result and conclusions.

2. Related Work

Pose estimation and tracking of the human figure has attracted substantial research interest over the past decade. Baumberg and Hogg [2] track the outline of a moving body using a flexible shape model. Ju et al. [12] proposed a parameterized motion model for tracking body parts, while shifting the focus of tracking from edges to the intensity pattern created by each body part in the image plane. Cham and Rehg [3] presented a probabilistic multiple-hypothesis framework for articulated figure tracking. They describe the

Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)
1063-6919/05 $20.00 © 2005 IEEE

figure by using Scaled Prismatic Models (SPM) and track the modes in the state pdf using a combination of parametric and sampling methods. In a follow-up work, Pavlovic et al. [14] addressed the problem of learning dynamics from training data in a switching Bayesian framework, focusing again on parametric models. Despite the use of dynamic priors, the inability to robustly handle self-occlusions presented a serious drawback of the former approaches. Methods that include explicit occlusion models based on layered representations of 2D images have been proposed in the past by, e.g., Jojic and Frey [10]. However, such models have typically not been used for articulated objects and often rely on non-parametric pose models.

The use of 3D figure models for pose estimation and tracking arose as a viable, albeit computationally more expensive, alternative that handles self-occlusions and resolves some of the ambiguities resulting from 3D singularities. Most of 3D human figure tracking approaches utilize sampling-based methods to handle multiple hypotheses and the non-linearities of the articulated and measurement models. However, the sampling approaches often suffer from the high dimensionality of the state space and the sample impoverishment problems. Annealed sampling by Deutscher et al. [6], importance sampling by Sidenbladh et al. [17], and partitioned sampling by MacCormick and Isard [13] were attempts to handle such problems. In the 2D case, Sminchisescu and Triggs [18] proposed a hybrid parametric method to alleviate the problems induced by the particle representation. Recently, alternative discriminative approaches to 3D pose estimation and tracking [20, 1] have shown initial promise, but remain in their infancy.

3. Articulated layers model

We present the formalism of the articulated layers model using an example of the human figure shown in Figure 1. In this model each articulated link is assigned to one image layer. The layers are ordered according to their depth in the 3D scene. Constraints imposed on the links across the layers guarantee the consistency of the articulated structure, e.g., layer 0 is above layer 1 and the parts in the two layers are directly linked.

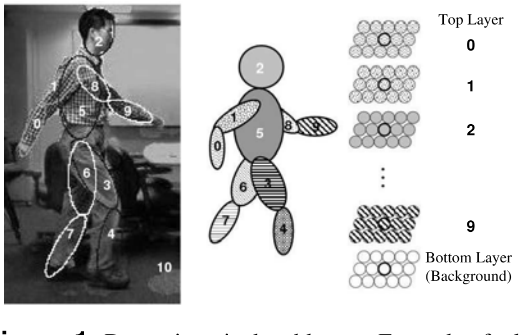

Figure 1. Dynamic articulated layers. Example of a layered articulated structure.

The model consists of four submodels whose graphical representations are shown in Figures 2-4. Similar models have been employed in other related Bayesian modeling approaches, e.g.,[10, 20]. We next discuss each of the four submodels.

3.1. Kinematic model

The first component is the kinematic model at time t, shown in Figure 2. We characterize the articulated model in terms of link joint angles $a_{i,t}$ and lengths $l_{i,t}$, with $i = \{1, ..., N\}$ the link index. For convenience of this presentation, unless we refer to the dynamic model, we often drop the time index t. We assume a prior distribution of link lengths, λ_i. a_i and l_i together determine the joint angle positions x_i in the image coordinate space, starting from the root position x_0. We assume that this dependency is deterministic[1] given the joint angles and the link lengths:

$$x_i = \mathbf{R}(a_i + a_{i-1} ... + a_1) \begin{bmatrix} l_i \\ 0 \end{bmatrix} + x_{i-1} \quad (1)$$

$$\mathbf{R}(a) = \begin{bmatrix} \cos(a) & -\sin(a) \\ \sin(a) & \cos(a) \end{bmatrix}. \quad (2)$$

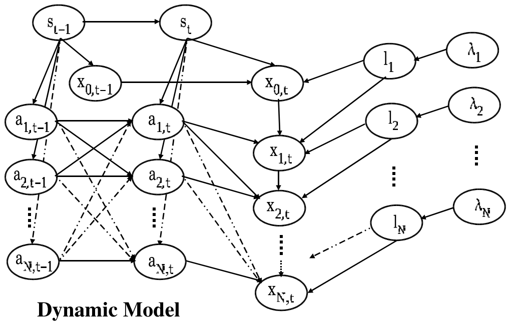

Figure 2. Dynamic and kinematic models.

3.2. Part pose and appearance models

We distinguish each link i from the corresponding articulated part i. The articulated part i is characterized by its pose and appearance, as depicted in Figure 3. The pose of part i is determined by its relative position y_i in the local coordinate system of the i-th link, relative orientation b_i wrt link i, and size d_i. The pose, in turn, has prior parameters γ_i, β_i, and δ_i. Together with the link position and orientation from the kinematic model, the relative pose induces the absolute part pose Y_i, B_i, d_i:

$$B_i = \sum_{j=0}^{i} a_j + b_i \quad (3)$$

[1] This assumption is not necessary. However, it simplifies the overall model without the loss of generality.

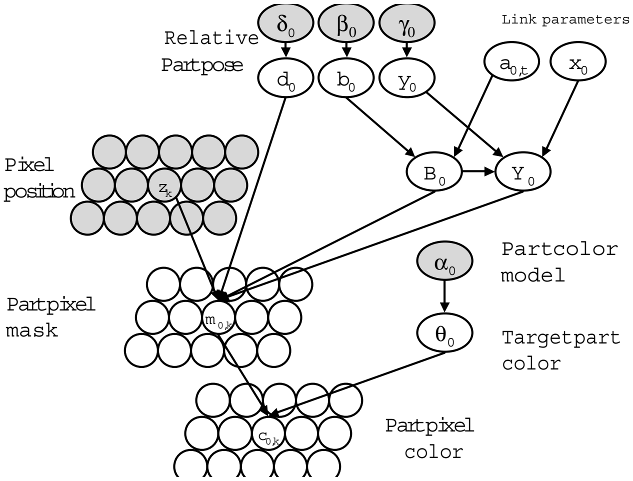

Figure 3. Part pose and appearance model.

$$Y_i \quad = \quad x_i + \mathbf{R}(B_i) \begin{bmatrix} y_i \\ 0 \end{bmatrix}. \tag{4}$$

In our models we commonly make the assumption that the relative part position $y_0 \doteq l_0/2$, i.e., the part is centered on link i. Furthermore, it is convenient to assume that the link is parallel with its respective part, $b_0 \doteq 0$.

The part's pose determines its appearance through the mask (region of interest) $\mathbf{m_i} \in \{0,1\}^{N_x \times N_y}$. This mask assigns image pixels k with known position z_k in the image coordinates to part i:

$$\begin{aligned} &P(m_{i,k} = 1 | Y_i, B_i, d_i, z_k) \sim \\ &\qquad \rho\left(|z_k - Y_i|' \mathbf{Q}^{-1}(B_i) |z_k - Y_i|\right), \end{aligned} \tag{5}$$

$$\mathbf{Q}(B_i) = \mathbf{R}(B_i) \begin{bmatrix} d_{i,x}^2 & 0 \\ 0 & d_{i,y}^2 \end{bmatrix} \mathbf{R}(B_i)'. \tag{6}$$

ρ is a robust kernel function with center Y_i and scale $\mathbf{Q}(B_i)$. Hence, each part is represented as an elliptical region of size d_i. In our models we often used a truncated Gaussian kernel:

$$\rho(x) = \left\{ \begin{array}{cc} \exp(-\frac{1}{2}x), & x < 1 \\ 0, & \text{otherwise} \end{array} \right. . \tag{7}$$

Inside the masked region, $\mathbf{m}_i = 1$, the appearance of part i is determined by the distribution θ_i of the appearance features $c_{i,k}$. Possible features are color, edges, texture, etc. A critical assumption we make is that, given the pose of the part, *all features inside the part are homogeneous*, i.e., they do not depend on the position of pixels in the part. This assumption allows computationally efficient estimation of the part's pose from image features. We also assume that all appearance features are discretized, e.g., 256 levels of gray or a set of discrete RGB values. In that case

$$P(c_{i,k} | \mathbf{m}_{i,k} = 1, \theta_i) = \text{Mult}(\theta_i), \tag{8}$$

where $\text{Mult}(\theta_i)$ is the multinomial distribution with parameters θ_i. This distribution can, in principle, vary from image to image, with some prior parameters α_i (e.g., Dirichlet or a mixture of Dirichlet distributions.) Such formulation can accommodate adaptive distribution estimation, due to, e.g., scene illumination changes or shadowing.

The appearance model proposed above is a generalization of the commonly used robust region-based image models. For instance, color distributions of this type have been used as the target model in several other tracking algorithms as they exhibit robustness against non-rigidity, illumination changes, and partial occlusions [4, 5, 15]. Unlike the heterogeneous representation of appearance in, e.g., [10] (each pixel has a potentially different parametrization), our representation also allows a computationally efficient and robust inference of the part's pose from appearance (see Section 4).

The pose and appearance of the background layer are an exception to the previously described model. We assume a static background layer, $P(m_{N,k} = 1) \doteq 1$. Its appearance is modeled by a set of heterogeneous features whose distribution depends on the pixel position, $P(c_{N,k} | m_{N,k}) = \text{Mult}(\theta_{N,k})$. In practice, however, we can often first remove the background layer using standard background subtraction methods.

3.3. Layered model

The layered model is the key link between the articulated model and the formed monocular image. For simplicity, we place each articulated part on a separate layer. Layers are then ordered according to their visibility i.e., the depth of the corresponding 3D scene. An example of this is depicted in Figure 1.

We define the layered representation similar to [10]. The ordering of layers and their corresponding parts is denoted by the link index i. Note that, for simplicity of notation, we so far assumed that the same index denotes the ordering of the links in the articulated model (e.g., root-to-leaves)[2]. Hence, the link with index 0 lays in the top layer, while the last layer N models the background scene.

The observed monocular image $\mathbf{c}$ is formed as a linear combination of the part masks $\mathbf{m}_i$, layer visibility $\mathbf{v}_i$ and latent layer images $\mathbf{c}_i$:

$$c_k = \sum_{i=0}^{N} v_{i,k} m_{i,k} c_{i,k}. \tag{9}$$

The visibility image $\mathbf{v}_i$ of layer i is induced recursively from the part masks in layers above i,

$$v_{i,k} = \prod_{j=0}^{i-1} (1 - m_{j,k}). \tag{10}$$

[2] This, of course, is not true in general. e.g., not true for the human model in Figure 1.

For instance, pixel k in layer i will be visible if it is not masked by any part in the previous $i-1$ layers. This image formation process is depicted in Figure 4. We note that the layered representation we adopted differs from the one in [10]—our image masks are binary variables unlike the continuous alpha blending values of [10]. Despite this assumption, our image formation model allows mixing of pixels intensities coming from different layers by the virtue of uncertainties in layer models.

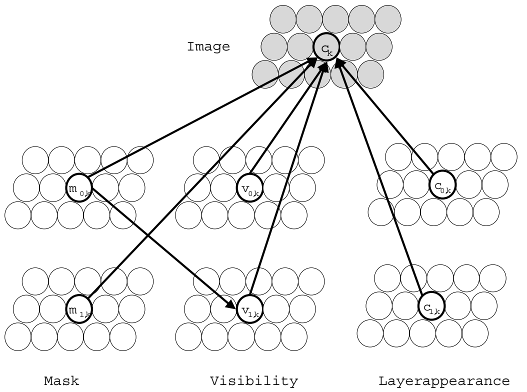

Figure 4. Image formation through layered representation.

3.4. Dynamic model

The final component of our model is the dynamic model. It links together articulated object poses across consecutive frames, as indicated in Figure 2. We assume a multiple hypothesis / switching linear dynamic model of [14]:

$$\mathbf{a}_t = \mathbf{A}(s_t)\mathbf{a}_{t-1} + \mathbf{w}, \tag{11}$$

where $\mathbf{a}_t$ is the vector of joint angles (and root link position) of the articulated model, and $\mathbf{w}$ is a white Gaussian noise. s_t denotes the id of the dynamic model at time t, among several possible models.

4. Pose estimation and tracking using layered articulated model

The ultimate goal of our approach is to estimate the pose of the tracked articulated object over a sequence of monocular image frames. Given the probabilistic formulation of our model this task can be accomplished using one of many inference methods for probabilistic models. In this section we derive a specific inference procedure that particularly suits our model.

We first focus on the problem of pose estimation in a single frame, given some prior pose estimate $P(\mathbf{a}_t)$. In the sequential pose estimation, this estimate typically comes from the predictions of the dynamic model and a pose estimate from the prior frame, $P(\mathbf{a}_t|\mathbf{c}_0, \ldots, \mathbf{c}_{t-1})$. Thus, in general, we need to infer $P(\mathbf{a}|\mathbf{c}, \text{prior on } \mathbf{a})$.

Unfortunately, the inference in the proposed model is intractable. Assuming, for simplicity, that the part dimensions $\mathbf{d}_j$, link lengths l_j and part appearance distributions θ_j are known, the intractability arises due to: (a) the existence of latent mask $\mathbf{m}_i$, visibility $\mathbf{v}_i$, and articulated joint positions $\mathbf{x}_j$, and (b) the nonlinear kinematic model. As expected, the latent variables introduce full coupling (dependency) between the observed image $\mathbf{c}_t$ and the object's pose.

To solve the pose estimation problem, we assume that the true posterior can be approximated by the product of two independent posteriors:

$$P(\mathbf{a}|\mathbf{c}) \approx \sum_{\mathbf{m}} Q_a(\mathbf{a}|\zeta_a) Q_m(\mathbf{m}|\mathbf{c}, \zeta_m). \tag{12}$$

$Q_a(\mathbf{a}|\zeta_a)$ is the posterior distribution of the object's pose, conditioned on a latent parameter set ζ_a. Similarly, $Q_m(\mathbf{m}|\mathbf{c})$ is the posterior distribution of the object's masks, conditioned on the observed image $\mathbf{c}$. The employed approximation is of a (structured) variational type [11].

To estimate the parts' masks we further assume that the posterior distribution Q_m is deterministic in the sense of

$$Q_m(\mathbf{m}|\mathbf{c}, \zeta_m) = \delta(\mathbf{m} - \mathbf{m}^*), \tag{13}$$

i.e., it is approximated with its most likely mode or a best mask estimate.

The latent variational parameter ζ_a as well as $\mathbf{m}^*$ can be found by minimizing the KL-divergence between the true and the approximate posteriors (see [11] for details). This yields a set of fixed point equations whose solutions are also the approximate posteriors for the pose and the masks:

```
input  : Image $\mathbf{c}$, prior on $\mathbf{a}$.
output : Posterior of $\mathbf{a}$, estimates of mask $\mathbf{m}$.
while !converged do
    Find modes of $Q_a(a|\zeta_a)$, ($\mathbf{a}^* = \arg\max_{\mathbf{a}} Q_a(\mathbf{a}|\zeta_a)$.
    Estimate $\mathbf{m}^* = \arg\max_{\mathbf{m}} Q_m(\mathbf{m}|\mathbf{c}, \zeta_m) = \arg\max_{\mathbf{m}} P(\mathbf{c}|\mathbf{m})P(\mathbf{m}|\mathbf{a}^*)$.
    Estimate $Q_a(\mathbf{a}|\mathbf{m}^*) \sim P(\mathbf{m}^*|\mathbf{a})P(\mathbf{a})$.
end
```

Algorithm 1: Pose and layer inference algorithm.

Intuitively, the algorithm first finds the optimal part masks given an estimate of the articulated object pose. It then uses the computed masks to refine the estimates of the articulated structure. The two steps are repeated until convergence. The algorithm is guaranteed to converge under fairly general conditions, as long as the two step are performed exactly [11].

4.1. Mask estimation

The solution to the mask estimation subproblem can be easily obtained by substituting the model definitions from Section 3 into the estimation equation in Algorithm 1. This, in turn, reduces to solving the following optimization problems for all pixels k in the image $\mathbf{c}$, starting from the top layer $i = 0$ down to the background $i = N$:

$$\begin{aligned} m^*_{i,k} = \arg \max_{m \in \{0,1\}} \\ [m\,(v_{i,k} \log \theta_i(c_k) + \log P(m_{i,k} = 1|\mathbf{a}, \mathbf{d})) + \\ (1-m)\,(v_{i,k}\hat{c}_{i+1:N} + \log P(m_{i,k} = 0|\mathbf{a}, \mathbf{d}))] \end{aligned} \quad (14)$$

As before, $v_{i,k}$ denotes the visibility of pixel k in layer i, (10). $\log \theta_i(c_k)$ is used to denote the likelihood of image feature c_k at pixel k under the model of layer i, θ_i. Finally, $\hat{c}_{i+1:N}$ is the average likelihood of the image pixel under the models below layer i, $\hat{c}_{i+1:N} = \sum_{j=i+1}^{N} \frac{v_{j,k}}{v_{i,k}} m_{j,k} \log \theta_j(c_k)$. The derivation above is similar to that of [10], however it differs in the underlying modeling assumptions.

4.2. Pose estimation

Given the estimates of the mask images $\mathbf{m}^*$ from Section 4.1 that remove the occlusion uncertainties and assign images to individual links, the estimates of the object's pose can now be obtained using one of several common methods such as the iterated extended Kalman filter (IEKF) [7]. We consider an alternative approach that robustly estimates the mode of the posterior pose.

In this approach, rather than maximizing the regularized likelihood score $P(m_{i,k}|\mathbf{a})P(\mathbf{a})$ over all poses $\mathbf{a}$, we maximize a Bhattacharyya error-based objective function between the target mask image and the image induced by the current mask position. This approach essentially extends the kernel-based tracking method of [5] to an articulated object setting.

We formulate the algorithm, without loss of generality, on the case of a two-link articulated object. In that case, the functional is

$$\begin{aligned} J(a_1, a_2) = \log P(a_1) + \tau_1 BT(\mathbf{m}_1|a_1, d_1) + \\ \tau_2 P(a_2|a_1) + \log BT(\mathbf{m}_2|a_2, d_2), \end{aligned} \quad (15)$$

where $BT(\mathbf{m}_i|a_i, d_i)$ denotes the Bhattacharyya distance between the mask image $\mathbf{m}_i$ and the current estimate of the kernel $\rho(\cdot)$'s scale and position, as specified in (5) and (7). This distance is defined as

$$\begin{aligned} BT(\mathbf{m}_i, |a_i, d_i) = \\ \sum_k \left(1 - \sqrt{m_{k,i} P(m_{i,k} = 1|Y_i, B_i, d_i, z_k)}\right) \end{aligned} \quad (16)$$

τ_1 and τ_2 are the precision weights that can be learned from data.

We maximize J using recursive forward-backward Newton-Raphson optimization steps. This approach slightly differs from the traditional M-estimate (mean-shift) search that typically only estimates the kernel positions, but is similar to robust position-scale estimators found in statistics literature, c.f.,e.g., [9]. We compute estimates of the variance of $\mathbf{a}$, needed for the dynamic predictions, based on the Hessians computed in the Newton-Raphson optimization step, following [3].

4.3. Dynamic prediction

To estimate the pose prior needed for the inference of pose in Section 4.2 we use standard dynamic prediction methods of linear dynamic systems. Because we characterize each pose estimate by its mode and variance, predicted pose estimates retain the same structure. If one is to use the switching formalism of [14] or multiple hypotheses of [3], the need arises for pruning of the exponential number of hypotheses. This could, in principle, be handled using a number of different techniques, such as mixture collapsing.

5. Experiments

We conducted two sets of experiments to study the performance and utility of the articulated layers tracking approach. The first set were the experiments on synthetic images, primarily aimed at the study of the proposed framework in Section 3 and the pose estimation algorithm of Section 4.2. In the second set, we applied the algorithm to tracking of various monocular image sequence of the human motion. The algorithm was implemented in MATLAB and the computation takes approximately 5 seconds/frame on Pentium 4 2.26GHz PC.

5.1. Synthetic data

An example of the pose estimation on a synthetic 64×64 image is shown in Figure 5. Background image was generated with a uniform probability over the entire gray scale (0 to 255). Part appearances had uniform distribution over 50 gray levels. The articulated object of known dimensions and part ordering had a probabilistic kinematic model with the joint angles normally distributed about certain means, with standard deviations of $\pi/40$. Initial pose estimates were randomly chosen, while minimally overlapping the true object. We chose the traditional (non-truncated) Gaussian kernel as the parts' mask model.

In 95% of the test cases, the algorithm converged within 4.2 pixels/joint of the true object pose. Figure 5(a) shows one run of the algorithm. In addition to the final pose estimates, we also considered the estimated states of the image masks $\mathbf{m}_i$, displayed in Figure 5(b). Images labeled as "Mask+Img" show the true estimates of the masks in their corresponding layers according to (14). We were also interested in seeing the benefits of combining the pose with

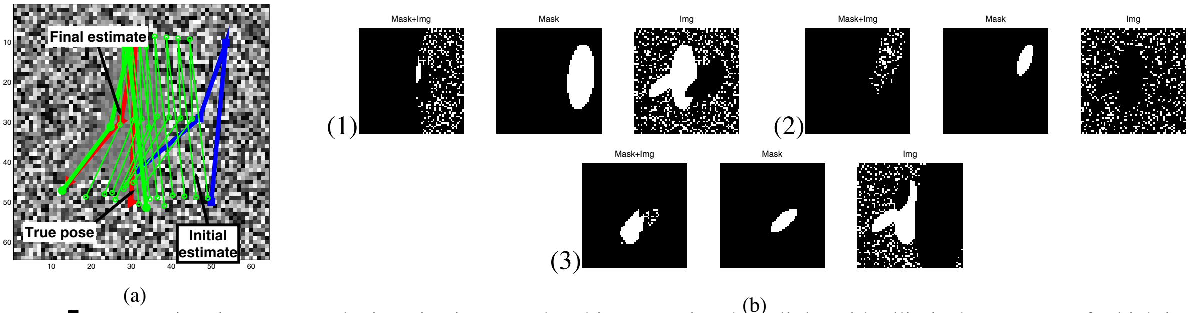

Figure 5. Pose estimation on a synthetic noisy image. The object contains three links with elliptical parts, one of which is fully occluded. The largest part is the top layer. Binary images represent results of mask estimation in three different contexts. See text for detailed explanation.

the image data, as in (14). Images "Mask" display mask predictions based on the current pose, using (5). On the other hand, images "Img" are the posterior estimates of the masks using the image data alone (obtained by setting the predicted mask terms, $P(m_{i,k} = 1|\mathbf{a})$ to 1/2.) One can clearly see the influence of the image/background noise and the out-of-layer parts on the estimates of masks using the image data alone. On the other hand, the mask predictions are, initially, far from the true part masks. Once the two estimates are integrated in the true mask posterior, the mask density peaks in the regions belonging to the respective object's part.

The algorithm always exhibited convergence, despite the approximations due to nonlinearity in the pose inference step, Section 4.2.

5.2. Real image sequences

The second set of experiments tested the utility of our approach for tracking articulated objects in real monocular image sequences. In particular, we focused on tracking and pose estimation of the human figure. The data used in these experiments was partially collected in-house; we also used the motion sequences made available under the HumanID project [8]. We employed a basic background subtraction procedure [3].

While the proposed framework, in principle, allows unsupervised learning of all model parameters, we initially estimated them from hand-labeled data. Furthermore, we assumed fixed and know ordering of the layers, as depicted in Figure 1. This assumption is reasonable as long as the object's 3D orientation does remain steady with respect to the imaging plane [4]. Initial pose of the figure was manually set in the first frame, as well as the scale which was kept constant for a given video sequence. We learned the appearance color model in HSV space from a set of random, segmented video frames. Finally, the dynamics within a single motion type (e.g., walking) were modeled using a linear model of the 1st or 2nd order (joint angles, angular velocities and accelerations, and corresponding linear motion states for the torso). Parameters of the model were estimated from training data.

Figure 6 illustrates the process of the pose estimation and layer separation inside a video frame. The final estimate of the pose is shown on the left. Images on the right depict the estimates of the part masks $\mathbf{m}_i$ and visibilities $\mathbf{v}_i$ for layers 6 through 9. For instance, the left-most images of the mask and visibility graphs correspond to layer 6, while the right most ones refer to layer 9.

Figures 7 and 8 show results of pose estimation and tracking on two video sequences of walking motion. To test the robustness of our approach neither of the instances used multiple hypotheses for the pose estimation or the dynamic predictions. Despite the common self-occlusions and shadowing, estimates of the best modes were sufficient to successfully track the pose over moderately long video sequences.

The pose estimation algorithm, similarly confirmed by our simulation studies in Section 5.1, showed robustness to poor dynamic model predictions. For instance, Figure 7 shows the dynamic model prediction (p) and pose estimation results (e) using the 1st order dynamic model. One can observe the relatively inadequate dynamic model predictions, compared to the 2nd order model employed in Figure 8. Even when the 2nd order dynamic prediction showed the lack of accuracy (bottom row of Figure 8), our alrorithm estimated the quite adequate pose (top row of Figure 8).

The most common failure modes occurred due to shadowing of the body parts, esp. the occluded arms and legs. Our present model does not attempt to model the changes in appearance due to shadowing. However, adaptation (on-line learning) of, e.g., color distribution, can be accomplished in

[3] The framework inherently includes a background subtraction step, given a known background model. This preprocessing step is strictly speaking unnecessary but, in practice, reduces the effects of complex backgrounds.

[4] Changes of orientation can, in principle, be handled using a set of view-specific models

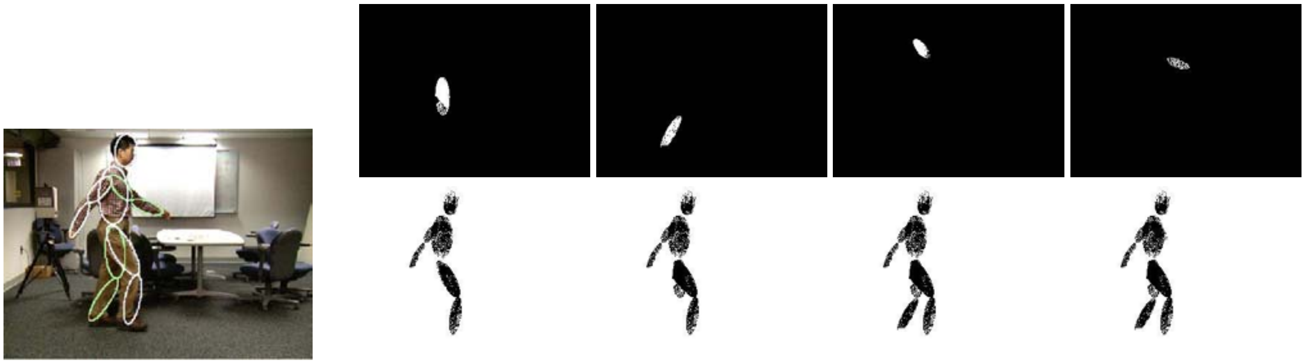

Figure 6. Inference of pose in image sequence. The human figure pose predicted by the articulated layers algorithm is shown on the left. Subsequent rows (dark background) show the estimated mask images and (white background) visibility images. White color corresponds to the 'on' or '1' state.

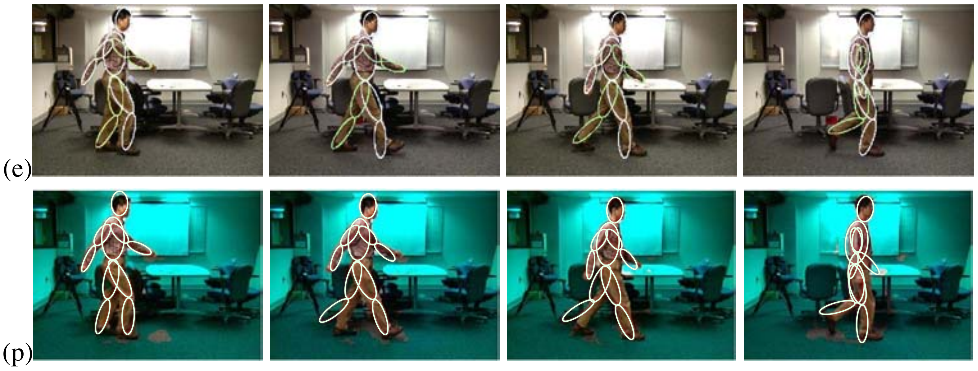

Figure 7. Tracking results for an in-house motion database sequence. (e) Pose estimates are marked using ellipses. (p) Dynamic model predictions are marked with ellipses on the background subtracted images (background is masked with blue color).

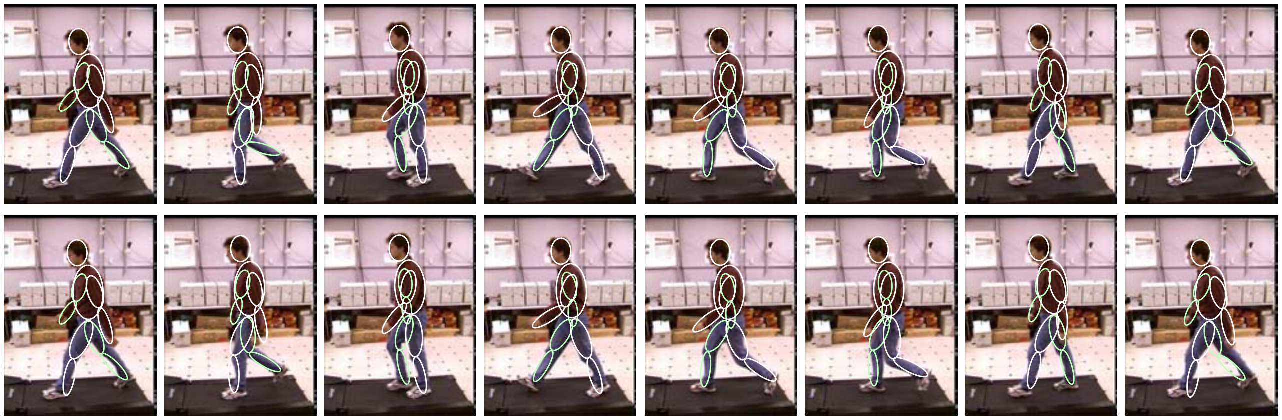

Figure 8. Tracking results for a CMU motion database sequence. Top row: pose estimates, bottom row: dynamic model predictions.

the same framework by inducing a stochastic dependency between the priors and current appearance parameters, α_i and θ_i. Similar adaptation of the other parameters (e.g., part sizes) could be performed in the same manner, at the additional computational cost.

The overall computational cost of the tracking algorithm is, on average, low. The layer mask estimation of Section 4.1 is proportional to the number of layers and the number of pixels inside each kernel's region of support. The estimation of the object's pose, described in Section 4.2, converges after a few iterations. Moreover, because the pose estimates is based on binary mask pixels $m_{i,k}$, not the image pixels c_k that belong to a much higher dimensional space, the BT computation is significantly simplified, compared to the traditional kernel-based methods [4, 5, 15].

6. Conclusions and future work

In this paper we proposed a probabilistic method for tracking articulated objects in monocular image sequence. To deal with the object self-occlusions, each link of the articulated object is assigned to a separate image layer. The appearance of object parts is represented using spatially homogeneous image features. The problem of pose estimation and tracking is then posed as an inference problem in this complex probabilistic model. We solve it using a recursive, robustified variational inference approach. Our preliminary results show the utility of this tracking approach on synthetic data as well as real video sequences. Despite the lack of full 3D object models, our method was able to successfully and robustly estimate and track 2D pose of the human figure across a number of image sequences.

Our future work proceeds in several directions. We will first consider additional image features as well as other inference methods for estimation of masks and part poses, e.g., true mask posterior estimates and IEKF for poses, and study their impact. Secondly, we will extend the formalism to include switching among several layered models, in order to handle continuous changes in the 3D object's orientation. Finally and most interestingly, we will consider approaches towards learning, unsupervised batch and on-line, of the model's parameters. This will include estimation of the ordering of layers in an object which is known to be an ill-posed problem.

References

[1] A. Agarwal and B. Triggs. Learning to track 3d human motion from silhouettes. In *International Conference on Machine Learning*, 2004.

[2] A. Baumberg and D. Hogg. Learning flexible models from image sequences. In *ECCV (1)*, pages 299–308, 1994.

[3] T.-J. Cham and J. M. Rehg. A multiple hypothesis approach to figure tracking. In *Proc. Conf. Computer Vision and Pattern Recognition*, pages 239–245, 1999.

[4] D. Comaniciu, V. Ramesh, and P. Meer. Real-time tracking of non-rigid object using mean shift. In *Proc. Conf. Computer Vision and Pattern Recognition*, volume II, pages 142–149, 2000.

[5] D. Comaniciu, V. Ramesh, and P. Meer. Kernel-based object tracking. *Pattern Analysis and Machine Intelligence*, 25(5):564–575, May 2003.

[6] J. Deutscher, A. Blake, and I. Reid. Articulated body motion capture by annealled particle filtering. In *CVPR*, volume II, pages 126–133, 2000.

[7] A. Gelb, editor. *Applied optimal estimation*. MIT Press, 1974.

[8] http://www.hid.ri.cmu.edu/Hid/databases.html.

[9] P. J. Huber. *Robust statistics*. Wiley, 1981.

[10] N. Jojic and B. Frey. Learning flexible sprites in video layers. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2001.

[11] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. In *Learning in graphical models*. 1998.

[12] S. X. Ju, M. J. Black, and Y. Yacoob. Cardboard people: A parameterized model of articulated motion. In *International Conference on Automatic Face and Gesture Recognition*, pages 38–44, Killington, Vermont, 1996.

[13] J. MacCormick and M. Isard. Partitioned sampling, articulated objects, and inference-quality hand tracker. In *ECCV*, 2001.

[14] V. Pavlovic, J. M. Rehg, T.-J. Cham, and K. P. Murphy. A dynamic bayesian network approach to figure tracking using learned dynamic models. In *ICCV*, pages 94–101, 1999.

[15] P. Perez, C. Hue, J. Vermaak, and M. Gangnet. Color-based probabilistic tracking. In *ECCV*, pages 661–675, 2002.

[16] R. Rosales and S. Sclaroff. Inferring body pose without tracking body parts. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 200.

[17] H. Sidenbladh, M. J. Black, and D. J. Fleet. Stochastic tracking of 3d human figures using 2d image motion. In *ECCV (2)*, pages 702–718, 2000.

[18] C. Sminchisescu and B. Triggs. Covariance scaled sampling for monocular 3d body tracking. In *Proc. International Conference on Computer Vision and Pattern Recognition*, 2001.

[19] C. Sminchisescu and B. Triggs. Kinematic jump processes for monocular 3d human tracking. In *International Conference on Computer Vision & Pattern Recognition*, pages I 69–76, June 2003.

[20] L. Taycher and T. Darrell. Bayesian articulated tracking using single frame pose sampling. In *Proc. 3rd Int'l Workshop on Statistical and Computational Theories of Vision*, 2003.

[21] S. Wachter and H.-H. Nagel. Tracking persons in monocular image sequences. *Comput. Vis. Image Underst.*, 74(3):174–192, 1999.