Articulated Multi-Body Tracking under Egomotion

S. Gammeter[1], A. Ess[1], T. Jäggli[1], K. Schindler[1], B. Leibe[1,2], and L. Van Gool[1,3]

ETH Zürich[1] RWTH Aachen[2] KU Leuven, IBBT[3]
{stephaga|aess|jaeggli|schindler|leibe}@vision.ee.ethz.ch

Abstract. In this paper, we address the problem of 3D articulated multi-person tracking in busy street scenes from a moving, human-level observer. In order to handle the complexity of multi-person interactions, we propose to pursue a two-stage strategy. A multi-body detection-based tracker first analyzes the scene and recovers individual pedestrian trajectories, bridging sensor gaps and resolving temporary occlusions. A specialized articulated tracker is then applied to each recovered pedestrian trajectory in parallel to estimate the tracked person's precise body pose over time. This articulated tracker is implemented in a Gaussian Process framework and operates on global pedestrian silhouettes using a learned statistical representation of human body dynamics. We interface the two tracking levels through a guided segmentation stage, which combines traditional bottom-up cues with top-down information from a human detector and the articulated tracker's shape prediction. We show the proposed approach's viability and demonstrate its performance for articulated multi-person tracking on several challenging video sequences of a busy inner-city scenario.

1 Introduction

Humans have truly amazing scene understanding capabilities. When walking in a busy street on our daily shopping tours, we are effortlessly aware of our surroundings; we can recognize other people visually, follow their tracks through crowded situations, and interpret their body poses from appearance cues. Computer vision is still a considerable way from this goal. While there has been a long history of research in articulated tracking (a good overview can be found in [8]), a vast majority of those papers focuses on recovering body poses of single persons in simpler environments [4, 6, 29, 22, 25, 26] (notable exceptions include [2, 15, 17, 20, 21, 31]). Although several approaches have demonstrated body pose estimation in static surveillance scenarios [15, 31], none of those systems addresses the more challenging task of articulated tracking in unconstrained, busy street scenes, where many people overlap and partially occlude each other and where the camera itself may undergo egomotion.

This is by no means a coincidence. Articulated tracking under such conditions is extremely hard, and many factors contribute to this difficulty. Even when only tracking a single person, pose estimation and data association between frames contain significant challenges. Several state-of-the-art approaches build up articulated models from local parts in a bottom-up fashion [20, 23]. Those approaches can easily get confused by the abundance of human limbs that are visible in busy street scenes. Other approaches rely on global shape, which is difficult to extract in crowded situations due to clutter, overlap and occlusion, especially when the camera itself is moving [11].

Fig. 1. *An example for the articulated multi-person tracking scenarios considered in this paper. We address this task by first applying a robust multi-body tracker to handle the data association problem and identify individual tracks. An articulated tracker is then applied to each single-person track independently to infer precise body poses, which are in turn fed back to improve the observation model. As can be seen from our results, this procedure allows robust performance despite the presence of multiple people, temporary occlusions, scale changes, and camera motion.*

When trying to track the articulations of multiple persons at the same time, additional difficulties arise from those persons' interactions. While algorithms that support multiple hypotheses can in principle deal with several people (see *e.g.* [20]), they typically do not explicitly distinguish between competing pose hypotheses for a single person in the image and different persons that are simultaneously visible. Also, relations between different subjects, such as temporary occlusion, cannot be modeled that way. A straightforward extension of a probabilistic inference algorithm to multiple subjects with occlusion reasoning requires a joint representation for the state space of multiple subjects [10], leading to an exponential increase in computational complexity.

In this paper, we propose an approach to overcome those difficulties in a system's context. The key insight behind this work is that it is not necessary to handle the complexity of multi-person interactions at the level of articulated tracking. Instead, we propose to carry out the global occlusion and multi-object reasoning on a coarser level and to only perform a more detailed articulated analysis on the output trajectories of the higher-level multi-body tracker (see Fig. 1). This allows us to also impart the articulated tracker with important information from trajectory analysis, such as a person's 3D walking direction, speed, and the knowledge when a trajectory is occluded. However, even a sophisticated multi-body tracker cannot solve the entire problem. Data association remains a challenging task: especially when multiple persons are walking close to each other, their limbs are often hard to distinguish. We address this issue by providing the articulated trackers with a guided segmentation that incorporates top-down knowledge from human detection. Together with a dynamic shape prediction from tracking, this observation model provides sufficiently precise measurements to support articulated multi-body tracking in very challenging street scenes.

In detail, this paper makes the following contributions. 1) We propose a combination of multi-body and articulated tracking for robust, multi-person 3D body pose estimation in inner-city scenes of realistic complexity. 2) We show how this general principle can be implemented in combination with a Gaussian Process (GP) articulated tracking approach based on global pedestrian silhouettes. This GP model incorporates learned prior knowledge of human shape and dynamics in order to capture the essence of human walking cycles. As regular GP training with sizable training sets is computationally very expensive, we propose several extensions to make learning tractable. 3) We augment the articulated tracker with a guided top-down/bottom-up segmentation procedure in order to reliably extract pedestrian silhouettes in busy scenes and under significant camera egomotion. 4) Finally, we demonstrate that our proposed system achieves robust performance in very challenging sequences.

The paper is structured as follows. The next section discusses related work. Section 3 gives an overview of our proposed system. The following three sections then present its different components in detail: the multi-body tracker (Sec. 4), our Gaussian Process articulated tracking approach (Sec. 5), and the guided top-down/bottom-up segmentation (Sec. 6). Finally, section 7 presents experimental results.

2 Related Work

The main challenges of 3D articulated tracking are the high-dimensional search spaces of body poses, multi-modal posterior distributions, and the fact that the images do not provide all the necessary information due to their 2D nature. Using multiple cameras and a controlled environment, ambiguities can be limited, and accurate 3D tracking results can be obtained [22, 4, 26]. We focus on realistic scenarios with noise and occlusions, where the scene is observed by a single camera or small-baseline stereo setup.

Many existing articulated tracking approaches can either be described as model-based generative top-down methods [25, 6], or part-based bottom-up approaches [20, 23] (see [8, 18] for a comprehensive survey). The latter typically only allow for pairwise constraints between neighboring body parts in a graphical model of the human body. In order to infer 3D body poses from monocular or binocular image sequences, more powerful holistic prior models of possible 3D poses have been learned in [25, 29].

More recently, several approaches have been proposed that learn the statistical properties of human body motion *and* the relationship between body poses and their image appearance. They rely on machine learning techniques such as kernel regressors or dimensionality reduction and can be divided into discriminative (*e.g.* [1, 27]) and generative (*e.g.* [14, 11, 19]) methods. While discriminative approaches lead to more direct inference algorithms, they have to deal explicitly with ambiguities of the one-to-many discriminative mapping. Furthermore, they assume that the subject's 2D image location is known beforehand, which is not a trivial task for challenging multi-person scenarios such as the ones considered here. Generative approaches, on the other hand, suffer from the high dimensionality of the body pose space, which is a problem for both the learning and the generative tracking algorithms. Their performance can however be improved by a suitable dimensionality reduction. [14, 11] first learn such a low-dimensional pose representation and then model the mappings into the pose and appearance spaces, as well

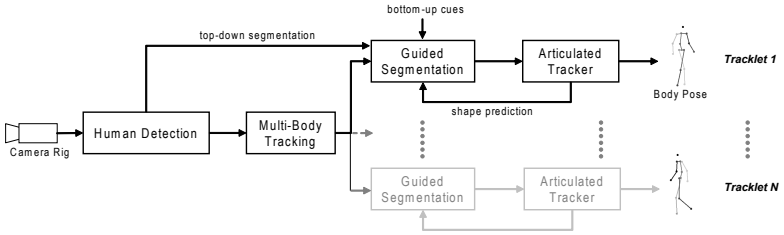

Fig. 2. Overview of the system.

as the pose dynamics, using kernel regressors. [19] proposes an integrated formulation that obtains a dimensionality reduction in a Gaussian Process framework by estimating a low-dimensional latent space which simultaneously maps into the pose and appearance spaces. In this paper, we follow a similar line, but explicitly take into account also the dynamics, which prove to be very important for our application.

While most articulated tracking approaches consider only single persons, several methods have also been proposed for multi-person scenarios. In [17], multiple independent articulated trackers are initialized manually on different persons. [21] also automates this initialization stage by detecting stylized poses for 2D body pose estimation. Several approaches have demonstrated 3D body pose estimation in static surveillance scenarios [15, 31]. Most directly related to our approach, [31] also applies a multi-object tracker to identify individual trajectories and estimate each tracked person's body poses over time. However, their tracking approach relies on background modeling, and their pose estimation process is relatively simple, based on a coarse discretization of the pose space. In very recent work, [2] propose an articulated pedestrian detector as basis for articulated multi-person tracking. While in this approach the articulation can help solve the data association problem, it is currently restricted to side-views and performs tracking only in 2D. In contrast, 3D articulated multi-body tracking from a moving, human-level perspective still remains an open issue.

3 System Overview

Fig. 2 shows the schematic layout of our multi-body articulated tracking system. A small-baseline (40cm) calibrated stereo rig mounted on a mobile platform captures two image streams and passes them on to a human detection module. Based on the obtained bounding boxes and rough stereo depth information, a multi-body tracker (Sec. 4) finds consistent object trajectories in 3D. Each trajectory is then passed to a single-person articulated tracker (Sec. 5), which estimates the person's 3D articulation based on a learned statistical representation. The estimation is made robust by a guided segmentation stage (Sec. 6) that combines the pedestrian detector's top-down segmentation with bottom-up image cues and a shape prediction inferred from the current state of the articulated tracker. This results, for every frame of the sequence, in one body pose estimate per tracked person, located in 3D world coordinates.

While in our approach, stereo-based depth computation supports the multi-body tracker and contributes to finding the subject's silhouette (see Sec.6) by setting it apart

from the background, the accuracy of the depth information is limited by the small baseline between the cameras and does not allow for further disambiguation of the pose estimates (as would be possible in a true multi-camera setup [4, 26, 22]). The articulated pose estimation algorithm thus relies on image descriptors that are computed from the subject's silhouette. We do however take into account both image streams of the binocular sequences, which helps to alleviate problems that are caused by image noise; *i.e.* when one camera stream is temporarily corrupted by noise or occlusion, the algorithm can base its pose estimation on the second camera.

4 Multi-Body Tracking

In order to reliably handle the complex interactions between multiple objects, we first address the task of tracking multiple pedestrians *without* taking into account the articulations, effectively factorizing the state space into independent "tracklets" for each visible pedestrian.

We adopt the tracking-by-detection approach from [7], which incorporates detection, stereo depth, and visual odometry to allow robust mobile tracking. The multi-body tracker itself is not the subject of this paper; we only give a high-level description of its functionality here and refer to [7] for details.

As input, the multi-person tracker takes two video streams recorded with our small-baseline stereo rig. A global world coordinate frame and ground-plane are recovered using structure-from-motion and stereo depth. Pedestrians are detected at each time-step with an ISM detector [16] and are placed in this global frame. Based on the space-time detections, an over-complete set of trajectories is tracked with independent Kalman filters. The best subset of this pool of hypotheses is selected through a global optimization procedure which enforces physical exclusion constraints, resulting in a consistent scene interpretation. The tracker is able to automatically initialize new tracks (usually, after about 5 detections) and to recover temporarily lost trajectories, thus enabling the system to track through occlusions.

The output of the tracker is a trajectory for each pedestrian in 3D world coordinates (including the person's 3D orientation, velocity, and bounding box), as well as the information when the person was occluded. As the articulated tracker is currently only trained on walking people, objects below and above a certain speed threshold are discarded. The unoccluded parts of each remaining trajectory (the "tracklets") can be processed independently by the subsequent articulated tracking module, which would otherwise become intractable. We want to point out, however, that data association between those tracklets still remains a challenging problem, as the limbs of adjacent persons may easily get confused. Section 6 therefore introduces a guided segmentation, which combines top-down information from the human detector with bottom-up image cues and which considerably improves the observation process.

5 3D Articulated Tracking

In this section, we present our articulated tracking approach. It operates on the output of the multi-body tracker and is provided with 3D trajectories and walking directions

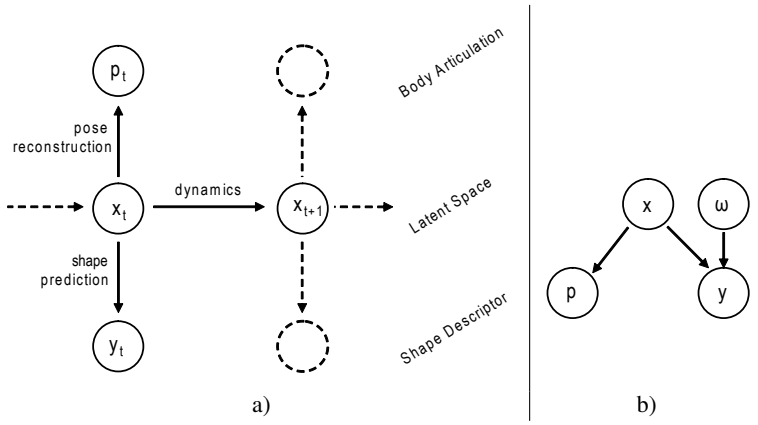

Fig. 3. Overview of the learned model. (a) Two slices of the temporal Markov chain. The arrows show the learned relationships between the variables. The low-dimensional pose representation ('Latent Space') is learned using LLE. (b) The prediction of the shape $\mathbf{y}$ depends on the low-dimensional body pose variable $\mathbf{x}$ and the orientation ω, while the body articulation $\mathbf{p}$ is only modeled as a function of $\mathbf{x}$.

of the individual pedestrians, where many ambiguities and temporary occlusions are already resolved and accounted for by the previous stage.

The articulated tracking algorithm is based on a learned statistical model of body motions and their appearance. This model follows a generative approach to capture the relationship between body pose and image appearance and is conceptually similar to [19]; here we additionally learn a dynamical model and propose extensions that make learning tractable for sizable training sets.

The training data consists of corresponding pairs of body articulations and shape descriptors (silhouettes). The body articulations are represented as a list of spatial 3D body part locations from 20 joints of the human skeleton, as shown in Fig. 1. The matching shape descriptors are vectors computed from a detected person's bounding box, where each entry indicates whether a certain pixel lies on the foreground or on the background. We currently use bounding boxes with a resolution of 45×50 pixels and apply PCA dimensionality reduction on the resulting 2250-dimensional shape descriptor.

The tracking algorithm operates in a low-dimensional representation of the body poses that is obtained by applying Locally Linear Embedding (LLE, [24]) on the data set of body articulations. We then model the *reconstruction* of the original representation of the articulations, the *prediction* of the corresponding human shape in image space, and the temporal evolution (*dynamics*) of the body poses over time using Gaussian Process regression [12]. This model is illustrated in Figure 3.

Pose reconstruction. Gaussian Processes define probability distributions over functions and can be used to model the regression between two variables, in our case the reconstruction from the low-dimensional pose space $\mathbf{X}$ to the original articulation representation $\mathbf{P}$. Given a covariance function $k_{rec}(\mathbf{x}_i, \mathbf{x}_j)$ and a set of training pairs, a posterior *pdf* over expected reconstructions $\mathbf{p}^*$ can be computed for any point $\mathbf{x}^*$ in the low-dimensional pose space. Training a GP regression model entails finding good

parameters β^{rec} of the covariance function (model selection). This can be done by maximizing the marginal likelihood of $\mathbf{P}$ with respect to the covariance parameters β^{rec}

$$P(\mathbf{P}|\mathbf{X}, \beta^{rec}) = (2\pi)^{-\frac{d_p N}{2}} |\mathbf{K}_{rec}|^{-\frac{d_p}{2}} exp(-\frac{1}{2} tr(\mathbf{K}_{rec}^{-1} \mathbf{P}\mathbf{P}^T)) \; . \tag{1}$$

Here, $\mathbf{X} \in \mathcal{R}^{N \times d_x}$ and $\mathbf{P} \in \mathcal{R}^{N \times d_p}$ are matrices containing the training data, N is the number of observations, and d_x and d_p are the respective dimensionalities of the appearance and pose data. The covariance matrix $\mathbf{K}_{rec} \in \mathcal{R}^{N \times N}$ is a function of the data $\mathbf{X}$ and the parameters β^{rec} of the covariance function, with elements $\mathbf{K}_{rec}^{i,j} = k_{rec}(\mathbf{x}_i, \mathbf{x}_j)$. We use standard squared exponential covariance functions with independent noise.

$$k_{rec}(\mathbf{x}_i, \mathbf{x}_j) = \beta_1^{rec} \exp\left(-\frac{\beta_2^{rec}}{2} ||\mathbf{x}_i - \mathbf{x}_j||^2\right) + \beta_3^{rec} \delta_{x_i, x_j} \; , \tag{2}$$

where $\beta^{rec} = \{\beta_1^{rec}, \beta_2^{rec}, \beta_3^{rec}\}$. The marginal likelihood (1) can then be optimized using numerical optimization methods such as scaled conjugate gradient.

Dynamics. In addition, our model is able to temporally predict future body poses according to a transition model $p(\mathbf{x}_{t+1}|\mathbf{x}_t)$. Similarly to (1), the marginal likelihood $P(\mathbf{X}|\beta^{dyn})$ is derived for the regression from $\mathbf{x}_t$ to $\mathbf{x}_{t+1}$ (see [30]), and optimized with respect to the parameters of the dynamics covariance function β^{dyn}.

Shape prediction. In contrast to the pose reconstruction, the shape prediction additionally depends on the orientation ω of the subject with respect to the observing camera (see Fig. 3b). In our training data, every body pose has a number of corresponding silhouettes, each viewed from a different angle. This results in NM training examples, where N is the number of poses and M the number of viewing directions in our training database. For the regression model, we thus have to optimize the marginal likelihood $P(\mathbf{Y}|\mathbf{\Omega}, \mathbf{X}, \beta^{app})$, where $\mathbf{\Omega}$ contains the viewing angles of the training shapes. Using a straightforward implementation, the complexity of the GP training algorithm scales with $(NM)^3$, since it involves the inversion of the covariance matrix $\mathbf{K}_{app} \in \mathcal{R}^{NM \times NM}$. This is impractical for the large datasets we use. We thus propose a covariance function that allows the covariance matrix to be written as a Kronecker tensor product, reducing the complexity to $O(N^3 + M^3)$ instead of the original $O((NM)^3)$. This can be done by defining the appearance covariance function as a product of a pose covariance function $k_{pose}(\mathbf{x}_i, \mathbf{x}_j)$ (*e.g.* squared exponential) and an orientation covariance function $k_{ori}(\omega_i, \omega_j)$,

$$k_{app}(\mathbf{x}_i, \omega_i; \mathbf{x}_j, \omega_j) = k_{pose}(\mathbf{x}_i, \mathbf{x}_j) k_{ori}(\omega_i, \omega_j) \; . \tag{3}$$

If for every pose $\mathbf{x} \in \mathbf{X} = \{\mathbf{x}_1 \ldots \mathbf{x}_N\}$ there are silhouettes for all possible viewing directions $\omega \in \mathbf{\Omega} = \{\omega_1 \ldots \omega_M\}$, then the appearance covariance matrix can be written as

$$\mathbf{K}_{app} = \mathbf{K}_{pose} \otimes \mathbf{K}_{ori} \tag{4}$$

Complexity can be further reduced by replacing the orientation covariance function with a delta function $k_{ori}(\omega_i, \omega_j) = \delta_{\omega_i, \omega_j}$. This makes sense during training of the GP regression, because the training samples only involve a number of discrete viewing directions $\omega \in \mathbf{\Omega}$. Once the regression parameters have been learned with this additional approximation, the orientation covariance function can then be replaced by one

with a larger support (*e.g.* a Von Mises distribution), in order to allow for interpolations between the discrete viewing directions $\omega \in \mathbf{\Omega}$.

Learning the embedding. The marginal likelihood of the entire learned model can now be written as

$$P(\mathbf{P}, \mathbf{Y}, \mathbf{X}|\mathbf{\Omega}, \beta^{rec}, \beta^{app}, \beta^{dyn}) = P(\mathbf{P}|\mathbf{X}, \beta^{rec}) P(\mathbf{Y}|\mathbf{\Omega}, \mathbf{X}, \beta^{app}) P(\mathbf{X}|\beta^{dyn}). \quad (5)$$

Rather than just optimizing the regressors, as done here, (5) could be optimized with respect to the latent positions $\mathbf{X}$ as well, where the LLE coordinates serve as an initialization, similarly to [19] . This would lead to a multi-set extension of the Gaussian Process Latent Variable Model [12] with separate covariance functions for each of the mappings. However, our experiments suggest that this does not improve the tracking results; they thus do not justify the increase in the number of parameters to be optimized from less than ten to several thousand.

Articulated Tracking. The articulated tracking algorithm operates on the output trajectories of the multi-body pedestrian tracker of Sec. 4, which delivers 2D image locations, scales, and orientations of the tracked persons. Its observations are automatically estimated pedestrian silhouettes, obtained through the guided segmentation procedure of Sec. 6. A particle filter serves as an overall framework, where at time t the body pose hypotheses $\mathbf{x}_t^i$ are propagated in the low-dimensional pose space according to the learned dynamical model. For each particle $\mathbf{x}_t^i$, a shape $\mathbf{y}_t^i$ can be predicted by taking into account the 3D track orientation ω_t, estimated by the multi-body tracker. The particles are then weighted with their image likelihoods, obtained by comparing the predicted shape to the actually observed shape $\mathbf{y}_t^{obs}$,

$$w^i \propto p(\mathbf{y}_t^{obs}|\omega_t, \mathbf{x}_t^i) = \mathcal{N}(\mathbf{y}_t^{obs}; \mu_t^i, \mathbf{\Sigma}_t^i) \,, \quad (6)$$

where μ_t^i and $\mathbf{\Sigma}_t^i$ are the mean and covariance matrix of the predicted shape.

Finally, once the particle filter has been run on all images of a tracklet, a Viterbi algorithm extracts a smooth and consistent trajectory through the particle set (note that this can in practice be approximated with a fixed temporal look-ahead). Again, the transition costs between neighboring states are based on the learned dynamical model. In order to account for variations in the framerate of the sequence and the walking speed of the subjects, this step additionally chooses between different scaling factors of the predicted velocities, *i.e.* accelerated and slowed-down variants of the dynamical model.

6 Guided Adaptive Segmentation

As an interface between the multi-body tracker and the articulated tracker, we are using a set of automatically estimated figure-ground segmentations for each tracked person. In the majority of previous works [19, 27, 31], silhouettes are assumed to be available and are in practice often obtained using background modeling. Since we are dealing with a moving camera setup, we cannot use this option. Instead, we propose to obtain the segmentations by fusing top-down cues (from the detector and the articulated tracker) with bottom-up image information (from color and stereo depth). Keeping in line with previous work by several authors [3, 5], the segmentation is formulated as

an energy minimization problem with respect to the foreground/background labelling $C = \{c_0, c_1, \ldots\}$ of all pixels.

$$E(C) = \sum_i R(c_i) + \lambda \sum_{i,j \in \mathcal{N}} B(c_i, c_j) \; . \tag{7}$$

In the above equation, $R(c_i)$ denotes the region term for a pixel with index i, which has a label c_i (figure/ground). $R(c_i)$ is based on the top-down segmentation map $\mathbf{f}$ of the detector and the shape prediction map $\boldsymbol{\pi} = \sum_j w_t^j \mu_t^j$ of the articulated tracker, where w_t^j is the weight of sample j and μ_t^j is its predicted shape from (6).

$$R(c_i) = -\log(P_\pi(c_i)\; P_f(c_i)) \tag{8}$$

Here, P_π and P_f are the probabilities of a certain label given the segmentation maps $\boldsymbol{\pi}$ and $\mathbf{f}$ from the articulated tracker and from the detector respectively:

$$P_\pi(c_i) = \begin{cases} \pi_i & \text{if } c_i = 1 \\ 1 - \pi_i & \text{if } c_i = 0 \end{cases} \; . \tag{9}$$

The boundary term $B(c_i, c_j)$ encodes the belief that region boundaries typically coincide with intensity and depth discontinuities. It is defined on the 4-neighborhood $\mathcal{N}$ and penalizes neighboring pixels with different labels but similar colors $\mathcal{I}_i$ and depths $\mathcal{D}_i$

$$B(c_i, c_j) = e^{-\frac{|\mathcal{I}_i - \mathcal{I}_j|^2}{2\sigma_c^2}} \; e^{-\frac{|\mathcal{D}_i - \mathcal{D}_j|^2}{2{\sigma_d}^2}} \; \delta_{c_i \neq c_j} \; . \tag{10}$$

The resulting cost function can be minimized efficiently using standard graph-cut methods [3], yielding the binary foreground mask $\mathbf{y}_t^{obs}$. Together with the bounding box position and motion direction from the multi-body tracker, this mask serves as the input for inference in the articulated tracker.

By ways of the expected appearance $\boldsymbol{\pi}$, we can incorporate prior knowledge about human shapes from the articulated tracker into the segmentation task, which can effectively complete partial segmentation maps $\mathbf{f}$. We can furthermore bridge missing detections by feeding back the tracker's expectation and combining it with bottom-up information. This is demonstrated in the sequence shown in Fig. 4: at first, the segmentation works well, giving a good initialization to the particle filter. In later frames, the detector fails due to missing contrast, but the prediction, along with the depth map, is good enough to obtain a usable segmentation.

Of course, care has to be taken not to reinforce erroneous feedback, which might lead to hallucinated walking cycles. Therefore, the influence of $\boldsymbol{\pi}$ is kept low as long as a detection is present and its weight is only increased when the trajectory contains holes. Even in those cases, however, the boundary terms usually restrict the segmentation well enough.

7 Results

Training. For training the articulated tracker, we recorded motion capture data (at 30 Hz) of 6 different people walking at speeds between 3 and 6 km/h. The resulting data set

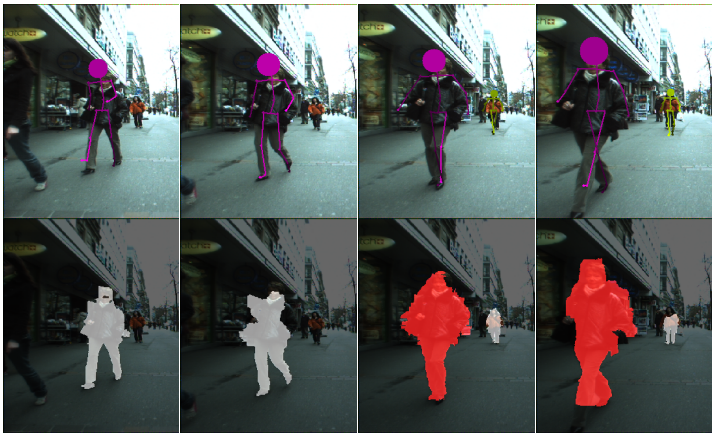

Fig. 4. *Original images (top) and sample segmentations obtained using our guided segmentation (bottom). If no detection is available, the articulated tracker's prediction is used in conjunction with bottom-up cues (red silhouettes).*

consists of slightly more than 2000 different body poses, each represented by 20 joint locations (*i.e.* 60 dimensions). For every body pose, silhouettes were rendered for 36 different viewing directions. A three-dimensional LLE of the body pose data serves as the low-dimensional pose space for the GP regression model. The marginal likelihood was optimized with scaled conjugate gradients using the FITC sparse approximation with 200 inducing variables [28, 13].

Results. We demonstrate our approach on 3 challenging video sequences showing real-world inner-city scenes. These videos were captured at about 13–14 fps from a mobile recording platform. Note that such a low framerate complicates the articulation reasoning considerably. Table 1 gives an overview over the sequences used in this paper. Video sequences are available on the following webpage: `http://www.vision.ee.ethz.ch/~stephaga/eccv2008/`

The first sequence (Fig. 5) was taken on a busy sidewalk. Even though the camera itself is standing still, traditional background subtraction would be difficult due to small camera shake, as well as passing trams and cars. While most people move sideways, they still occur at different depths and often have a slightly tilted trajectory, which we can account for by tracking directly in 3D. In the sequence's 454 frames, our system tracks 20 out of 23 people successfully with the multi-body tracker. One of the missed pedestrians runs too fast, and another one is at all times occluded by other persons. In addition to the 20 correct tracks, the system yields two additional tracks that contain errors due to wrongly estimated orientation or scale. Counting each person individually, this amounts to a total track length of 932 frames, where a detection is available in 86% of the cases. We visually inspected all the resulting segmentations and found that 55% of these are well-defined (meaning the entire person is covered) in at least one camera. For the individual cameras, only 41% were well-defined. This underlines the usefulness of a stereo system in such real-world scenarios with frequent occlusions. While these numbers might seem low, we would like to note that the articulated tracker can also operate if only parts of the body are segmented correctly (most importantly, the legs).

Seq.	# Frames	Pedestrians	Found by MBT
#1	454	23	20
#2	173	14	10
#3	242	21	17

Table 1. Sequences used for evaluating the proposed system. We report the number of (walking) persons, and the ones actually found by the multi-body tracker (MBT).

Based on these segmentations, the system tracks 74 walking cycles, 54 out of which were entirely correct. The remaining 20 cases mainly occur at the end of longer trajectories and are mostly due to multiple, consecutive bad segmentations or occlusions. Note, however, that the silhouettes generally did not contain enough information to unambiguously recover the arm positions, which additionally differed from our training examples since many people were carrying shopping bags or similar accessories. Example pose estimates of our system are shown in Fig. 5.

For the remaining sequences, we show qualitative results in Figs. 6 and 7. As the multi-body tracker takes care of the mapping between the world coordinate frame and the local articulated trackers, we can apply our system to scenes captured under significant egomotion. Fig. 6 shows an example of such a case, where people enter the scene from several directions and undergo large scale changes. As the multi-body tracker restricts the sampling for orientations, we can still get acceptable results on such data. A more challenging case is shown in Fig. 7. Here, the system has to cope with more extreme scale changes and people moving in many different directions, while following one person through a busy pedestrian zone. In particular movement parallel to the viewing direction is highly ambiguous; still the articulation is identified in most cases.

8 Conclusion

We have presented a system for articulated multi-body articulated tracking from a moving platform. Our approach achieves good results in challenging real-world scenarios by factorizing the problem into separate tasks of multi-body tracking under occlusion and articulated body pose estimation for individual trajectories. This formulation allows the articulated tracker to benefit from trajectory-level information about the tracked person's speed and walking direction, which considerably simplifies inference and renders the problem tractable. We have further presented a way to implement this idea with an articulated tracker based on Gaussian Processes and have shown how the framework can be applied under egomotion with the help of a guided top-down/bottom-up segmentation module. Experimental results confirm the viability of our proposed approach.

Currently, our method is restricted to learned articulations from known actions such as walking and running; in contrast to bottom-up approaches it cannot recover arbitrary body poses. We are exploring ways to replaces the global body models by semi global ones to mitigate this issue. In addition, the results of our estimation could be used to learn specialized color models for different body parts, which then support more general pose recovery [21]. The method for pose inference from [2] could also be a promising extension, since it is based on local appearance and may enable a more direct interplay between detector and tracker. In addition, we will enlarge the training set of learned gait dynamics in order to more densely cover the range of different walking styles and will

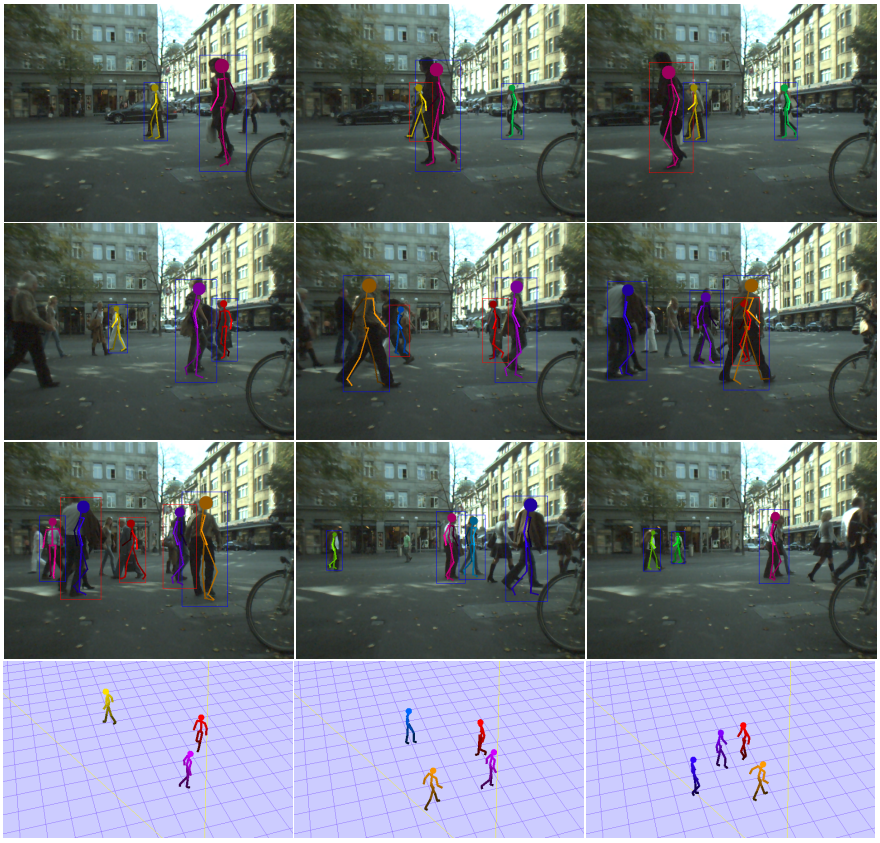

Fig. 5. *Articulated multi-person tracking results for test sequence #1. The last row shows a 3D visualization of the estimated world state in the three images of the second row (This figure is best viewed in color).*

include gait modifications when carrying luggage items. Finally, we plan to extend the feedback from body pose estimation to the multi-body tracker in order to also improve the detection model by incorporating gait dynamics, similar to [9].

Acknowledgements The authors thank Tobias Müller for developing the graph-cut-based segmentation. This project has been funded in parts by Toyota Motor Corporation/Toyota Motor Europe and the EU projects DIRAC (IST-027787) and HERMES (IST-027110).

References

1. A. Agarwal and B. Triggs. Recovering 3d human pose from monocular images. *PAMI*, 28(1):44–58, 2006.
2. M. Andriluka, S. Roth, and B. Schiele. People-tracking-by-detection and people-detection-by-tracking. In *CVPR'08*.
3. Y. Boykov and G. F. Lea. Graph cuts and efficient n-d image segmentation. *IJCV*, 70(2), 2006.
4. J. Carranza, C. Theobalt, M. Magnor, and H.-P. Seidel. Free-viewpoint video of human actors. In *SIGGRAPH'03*.

Fig. 6. *Articulated tracking results for test sequence #2. Note the robust articulation estimates of tracked pedestrians under scale changes and egomotion. (This figure is best viewed in color).*

5. D. Cremers, M. Rousson, and R. Deriche. A review of statistical approaches to level set segmentation: integrating color, texture, motion and shape. *IJCV*, 72, 2007.
6. J. Deutscher, A. Blake, and I. Reid. Articulated body motion capture by annealed particle filtering. In *CVPR'00*.
7. A. Ess, B. Leibe, K. Schindler, and L. van Gool. A mobile vision system for robust multi-person tracking. In *CVPR'08*.
8. D. Forsyth, O. Arikan, L. Ikemoto, J. O'Brien, and D. Ramanan. Computational studies of human motion: Part 1. *Foundations Trends Comp. Graphics Vision*, 1(2/3), 2006.
9. J. Giebel, D. Gavrila, and C. Schnörr. A bayesian framework for multi-cue 3d object tracking. In *ECCV'04*.
10. C. Hue, J.-P. L. Cadre, and P. Prez. Tracking multiple objects with particle filtering. *IEEE Trans. on Aerospace and Electronic Systems*, 38(3), 2002.
11. T. Jaeggli, E. Koller-Meier, and L. Van Gool. Learning generative models for monocular body pose estimation. In *ACCV'07*.
12. N. Lawrence. Probabilistic non-linear principal component analysis with gaussian process latent variable models. *JMLR*, 6, 2005.
13. N. D. Lawrence. Learning for larger datasets with the gaussian process latent variable model. In *Proc. Intern. Workshop on Artificial Intelligence and Statistics*, 2007.
14. C.-S. Lee and A. Elgammal. Modeling view and posture manifolds for tracking. In *ICCV'07*.
15. M.-W. Lee and R. Nevatia. Human pose tracking using multi-level structured models. In *ECCV'06*.
16. B. Leibe, A. Leonardis, and B. Schiele. Robust Object Detection with Interleaved Categorization and Segmentation. *IJCV*, 77(1–3), 2008.
17. J. Mitchelson and A. Hilton. Simultaneous pose estimation of multiple people using multiple-view cues with hierarchical sampling. In *BMVC'03*.

Fig. 7. *Articulated tracking results for test sequence #3. This sequence shows a very challenging scenario with considerable egomotion and many pedestrians entering the visible scene at various distances and from different directions. (This figure is best viewed in color).*

18. T. Moeslund, A. Hilton, and V. Krüger. A survey of advances in vision-based human motion capture and analysis. *CVIU*, 104(2), 2006.
19. R. Navaratnam, A. W. Fitzgibbon, and R. Cipolla. The joint manifold model for semi-supervised multi-valued regression. In *ICCV'07*.
20. D. Ramanan and D. Forsyth. Finding and tracking people from the bottom up. In *CVPR'03*.
21. D. Ramanan, D. Forsyth, and A. Zisserman. Strike a pose: Tracking people by finding stylized poses. In *CVPR'05*.
22. L. Ren, G. Shaknarovich, J. Hodgins, H. Pfister, and P. Viola. Learning silhouette features for control of human motion. *ACM Trans. Graphics*, 24(4), 2005.
23. X. Ren, A. Berg, and J. Malik. Recovering human body configurations using pairwise constraints between parts. In *ICCV'05*.
24. S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500), 2000.
25. H. Sidenbladh, M. Black, and D. Fleet. Stochastic tracking of 3d human figures using 2d image motion. In *ECCV'00*.
26. L. Sigal, S. Bhatia, S. Roth, M. Black, and M. Isard. Tracking loose-limbed people. In *CVPR'04*.
27. C. Sminchisescu, A. Kanaujia, Z. Li, and D. Metaxas. Discriminative density propagation for 3d human motion estimation. In *CVPR'05*.
28. E. Snelson and Z. Ghahramani. Sparse Gaussian processes using pseudo-inputs. In *NIPS'06*.
29. R. Urtasun, D. Fleet, A. Hertzmann, and P. Fua. Priors for people tracking from small training sets. In *ICCV'05*.
30. J. Wang, D. Fleet, and A. Hertzmann. Gaussian process dynamical models. In *NIPS'06*.
31. T. Zhao and R. Nevatia. Tracking multiple humans in complex situations. *PAMI*, 26(9), 2004.