Monocular 3D Pose Estimation and Tracking by Detection

Mykhaylo Andriluka[1,2] Stefan Roth[1] Bernt Schiele[1,2]
[1]Department of Computer Science, TU Darmstadt [2]MPI Informatics, Saarbrücken

Abstract

Automatic recovery of 3D human pose from monocular image sequences is a challenging and important research topic with numerous applications. Although current methods are able to recover 3D pose for a single person in controlled environments, they are severely challenged by real-world scenarios, such as crowded street scenes. To address this problem, we propose a three-stage process building on a number of recent advances. The first stage obtains an initial estimate of the 2D articulation and viewpoint of the person from single frames. The second stage allows early data association across frames based on tracking-by-detection. These two stages successfully accumulate the available 2D image evidence into robust estimates of 2D limb positions over short image sequences (= tracklets). The third and final stage uses those tracklet-based estimates as robust image observations to reliably recover 3D pose. We demonstrate state-of-the-art performance on the HumanEva II benchmark, and also show the applicability of our approach to articulated 3D tracking in realistic street conditions.

1. Introduction

This work addresses the challenging problem of 3D pose estimation and tracking of multiple people in cluttered scenes using a monocular, potentially moving camera. This is an important problem with many applications including video indexing, automotive safety, or surveillance. There are multiple challenges that contribute to the difficulty of this problem and need to be addressed simultaneously. Probably the most important challenge in articulated 3D tracking is the inherent ambiguity of 3D pose from monocular image evidence. This is particularly true for cluttered real-world scenes with multiple people that are often partially or even fully occluded for longer periods of time. Another important challenge, even for 2D pose recovery, is the complexity of human articulation and appearance. Additionally, complex and dynamically changing backgrounds of realistic scenes complicate data association across multiple frames. While many of these challenges have been addressed individually, we are not aware of any work that has addressed all of them simultaneously using a monocular, potentially moving camera.

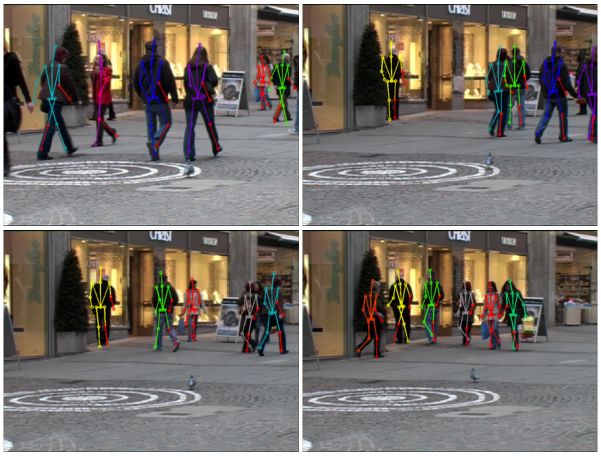


Figure 1. **Example results**: 3D tracking in a challenging scene.

The goal of this paper is to contribute a sound Bayesian formulation to address this challenging problem. To that end we build on some of the most powerful approaches proposed for people detection and tracking in the literature. In three successive stages we accumulate the available 2D image evidence to enable robust 3D pose recovery.

Overview of the approach. Ultimately, our goal is to estimate the 3D pose Q_m of each person in all frames m of a sequence of length M, given the image evidence $\mathcal{E}_{1:M}$ in all frames. To that end, we define a posterior distribution over pose parameters given the evidence:

$$p(Q_{1:M}|\mathcal{E}_{1:M}) \propto p(\mathcal{E}_{1:M}|Q_{1:M})p(Q_{1:M}). \tag{1}$$

Here, $Q_{1:M}$ denotes the 3D pose parameters over the entire sequence. Clearly, a key difficulty is that the posterior in Eq. 1 has many local optima as the estimation of 3D poses is highly ambiguous given monocular images. To address this problem this paper proposes a new three-stage approach sequentially reducing the ambiguity in 3D pose recovery.

Before giving an overview of the three-stage process, let us define the observation likelihood $p(\mathcal{E}_{1:M}|Q_{1:M})$. We assume conditional independence of the evidence in each frame given the 3D pose parameters Q_m. The likelihood thus factorizes into single-frame likelihoods:

$$p(\mathcal{E}_{1:M}|Q_{1:M}) = \prod_{m=1}^{M} p(\mathcal{E}_m|Q_m). \tag{2}$$

In this paper, the evidence in each frame is represented by the estimate of the person's 2D viewpoint w.r.t. the camera and the posterior distribution of the 2D positions and orientations of body parts. To estimate these reliably from single frames, the *first stage* (Sec. 2) builds on a recently proposed part-based people detection and pose estimation framework based on discriminative part detectors [3].

To accumulate further 2D image evidence, the *second stage* (Sec. 3) extracts people tracklets from a small number of consecutive frames using a 2D-tracking-by-detection approach. Here, the output of the first stage is refined in the sense that we obtain more reliable 2D detections of the people's body parts as well as more robust viewpoint estimates.

The *third stage* (Sec. 4) then uses the image evidence accumulated in the previous two stages to recover 3D pose. As described later, we model the temporal prior $p(Q_{1:M})$ over 3D poses as a hierarchical Gaussian process latent variable model (hGPLVM) [17]. We combine this with a hidden Markov model (HMM) that allows to extend the people-tracklets, which cover only a small number of frames at a time, to possibly longer 3D people-tracks. Note that our 3D model is assumed to generate the bottom-up evidence from 2D body models and thus constitutes a hybrid generative/discriminative approach (c.f. [26]).

The contributions of this paper are two-fold. The main contribution is a novel approach to human pose estimation, which combines 2D position, pose and viewpoint estimates into an evidence model for 3D tracking with a 3D motion prior, and is able to accurately estimate 3D poses of multiple people from monocular images in realistic street environments. The second contribution, which serves as a building block for 3D pose estimation, is a new pedestrian detection approach based on a combination of multiple part-based models. While the power of part-based models for people detection has already been demonstrated (*e.g.*, [3]), here we show that combining multiple part-based models leads to significant performance gains, and while improving over the state-of-the art in detection, also allows to estimate viewpoints of people in monocular images.

Related Work. Due to the difficulties involved in reliable 3D pose estimation, this task has often been considered in controlled laboratory settings [4, 6, 13, 24, 29], with solutions frequently relying on background subtraction and simple image evidence, such as silhouettes or edge-maps. In order to constrain the search in high-dimensional pose spaces these approaches often use multiple calibrated cameras [13], complex dynamical motion priors [29], or detailed body models [4]. Their combination allows to achieve impressive results [13, 15], similar in performance to commercial marker-based motion capture systems.

However, realistic street scenes do not satisfy many of the assumptions made by these systems. For such scenes multiple synchronized video streams are difficult to obtain, the appearance of people is significantly more complex, and robust extraction of evidence is challenged by frequent full and partial occlusions, clutter, and camera motion. In order to address these challenges, a number of methods leverage recent advances in people detection and either use detection for prefiltering and initialization [11, 14], or integrate detection, tracking and pose estimation within a single "tracking-by-detection" framework [2].

This paper builds on state-of-the-art people detection and 2D pose estimation and leverages recent work in this area [3, 7, 10, 20], which we combine with a dynamic motion prior [2, 28]. While [2] has shown to enable 2D pose estimation for people in sideviews, this paper goes beyond by estimating poses in 3D from multiple viewpoints. Compared to [14], we are able to estimate poses in monocular images, while their approach uses stereo. [11] proposes to combine detection and 3D pose estimation for monocular tracking, but relies on the ability to detect characteristic poses of people, which we do not require here.

Estimation of 3D poses from 2D body part positions was previously proposed in [25]. However, this approach was evaluated only in laboratory conditions for a single subject and it remains unclear how well it generalizes to more complex settings with multiple people as considered here.

There is also a lot of work on predicting 3D poses directly from image features using regression [1, 16, 27], classification [21], or a search over a database of exemplars [19, 22]. These methods typically require a large database of training examples to achieve good performance and are challenged by the high variability in appearance of people in realistic settings. In contrast, we represent the complex appearance of people using separate appearance models for each body part, which reduces the number of required training images and makes our representation more flexible.

2. Multiview People Detection in Single Frames

2D people detection and pose estimation serves as one of our key building blocks for 3D pose estimation and tracking. Our approach is driven by three major goals: (1) We want to take advantage of the recent developments in 2D people detection and pose estimation to define robust appearance models for 3D pose estimation and tracking; (2) we aim to reduce the search space of possible 3D poses by taking advantage of inferred 2D poses; and (3) we want to extract the viewpoint from which people are visible to reduce the inherent 2D-to-3D ambiguity. To that end we build on and extend a recent 2D people detection approach [3].

2.1. Basic pictorial structures model

Pictorial structures [8] represent objects, such as people, as a flexible configuration of N different parts $L_m = \{\mathbf{l}_{m0}, \mathbf{l}_{m1}, \ldots, \mathbf{l}_{mN}\}$. m denotes the current frame of

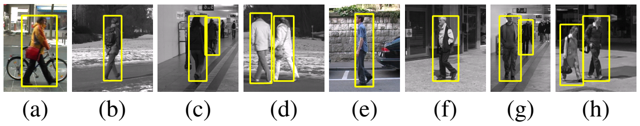


Figure 2. Training samples shown for each viewpoint: (a) right, (b) r.-back, (c) b., (d) left-b., (e) l., (f) l.-front, (g) f., (h) r.-f.

the sequence. The state of part i is given by $\mathbf{l}_{mi} = \{x_{mi}, y_{mi}, \theta_{mi}, s_{mi}\}$, where x_{mi} and y_{mi} denote its image position, θ_{mi} the absolute orientation, and s_{mi} the part scale. The posterior probability of the 2D part configuration L_m given the single frame image evidence D_m is given as

$$p(L_m|D_m) \propto p(D_m|L_m)p(L_m). \tag{3}$$

The prior on body configurations $p(L_m)$ has a tree structure and represents the kinematic dependencies between body parts. It factorizes into a unary term for the root part (here, the torso) and pairwise terms along the kinematic chains:

$$p(L_m) = p(\mathbf{l}_{m0}) \prod_{(i,j)\in K} p(\mathbf{l}_{mi}|\mathbf{l}_{mj}), \tag{4}$$

where K is the set of edges representing kinematic relationships between parts. $p(\mathbf{l}_{m0})$ is assumed to be uniform, and the pairwise terms are taken to be Gaussian in the transformed space of the joints between the adjacent parts [3, 8]. The likelihood term is assumed to factorize into a product of individual part likelihoods

$$p(D_m|L_m) = \prod_{i=0}^{N} p(\mathbf{d}_{mi}|\mathbf{l}_{mi}). \tag{5}$$

To define the part likelihood, we rely on the boosted part detectors from [3], which use the truncated output of an AdaBoost classifier [12] and a dense shape context representation [5, 18]. Our model is composed of 8 body parts: left/right lower and upper legs, torso, head and left/right upper and lower arms (later sideview detectors also use left/right feet for better performance).

Apart from its excellent performance in complex real world scenes [3, 7], the pictorial structures model also has the advantage that inference is both optimal and efficient due to the tree structure of the model. We perform sum-product belief propagation to compute the marginal posteriors of individual body parts, which can be computed efficiently using convolutions [8].

Data and evaluation. While the detector of [3] is in principle capable of detecting people from arbitrary views, its detection performance has only been evaluated on side views. To evaluate its suitability for our multiview setting, we collected a dataset of 1486 images for training, 248 for validation, and 248 for testing, which we carefully selected so that sufficiently many people are visible from all viewpoints. In addition to the persons' bounding boxes, we also annotated the viewpoint of all people in our dataset by assuming 8 evenly spaced viewpoints, each 45 degrees apart from each other (front/back, left/right, and diagonal front/back left/right). Fig. 2 shows example images from our training set, one for each viewpoint.

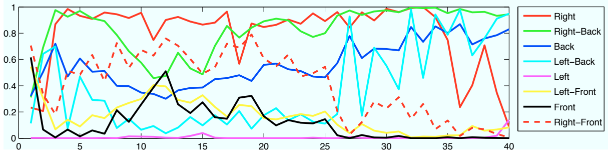


Figure 3. Calibrated output of the 8 viewpoint classifiers.

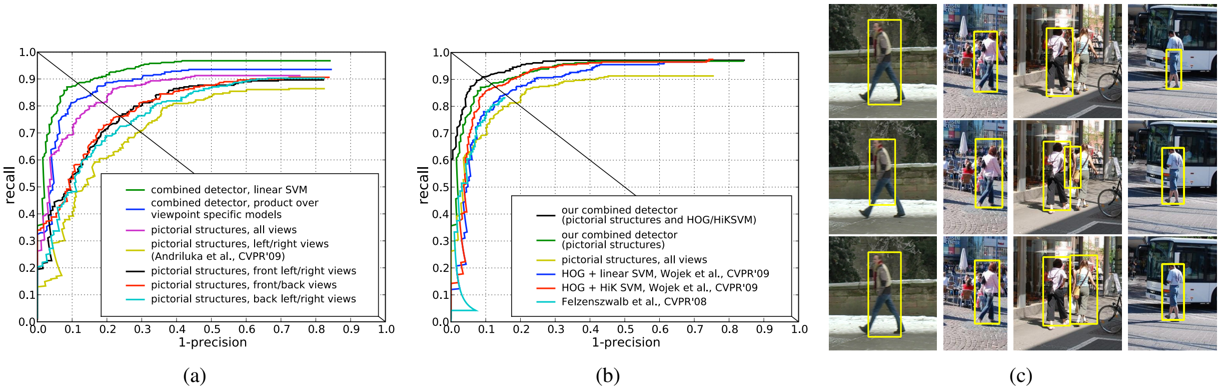
As expected and can be seen in Fig. 4(a), the detector trained on side-views as in [3] shows only modest performance levels on our multiview dataset. By retraining the model on our multiview training set we obtain a substantial performance gain, but still do not achieve the performance levels of monolithic, discriminative HOG-based detectors [30] or HOG-based detectors with parts [9] (see Fig. 4(b)). However, since we not only need to detect people, but also estimate their 2D pose, such monolithic or coarse part-based detectors are not appropriate for our task.

2.2. Multiview Extensions

To address this shortcoming, we develop an extended multiview detector that allows 2D pose estimation as well as viewpoint estimation. We train 8 viewpoint-specific detectors using our viewpoint-annotated multiview data. These viewpoint-specific detectors not only have the advantage that their kinematic prior is specific to each viewpoint, but also that part detectors are tuned to each view. We enrich this set of detectors with one generic detector trained on all views, as well as two side-view detectors as in [3] that additionally contain feet (which improves performance).

We explored two strategies for combining the output of this bank of detectors: (1) We simply add up the log-posterior of a person being at a particular image location as determined by the different detectors; and (2) we train a linear SVM using the 11-dimensional vector of the mean/variance-normalized detector outputs as features. The SVM detector was trained on the validation set of 248 images. Fig. 4(a) shows that the simple additive combination of viewpoint-specific detectors improved over the detection performance from each individual viewpoint-specific detector. It also outperforms the approach from [3].

Interestingly, the new SVM-based detector not only substantially improves performance, but also outperforms the current state-of-the-art in multiview people detection [9, 30]. As is shown in Fig. 4(b), the performance improves even further when we extend our bank of detectors with the HoG-based detector from [30]. While this is not the main focus of our work, this clearly shows the power of the first stage of our approach. Several example detections in Fig. 4(c) demonstrate the benefits of combining viewpoint-specific detectors.



(a) (b) (c)

Figure 4. Comparison between (a) viewpoint-specific models and combined model, and (b) comparison to state-of-the-art on the "MultiviewPeople" dataset; (c) sample detections obtained with the side-view detector of [3] (top), the generic detector trained on our multiview dataset (middle), and the proposed detector combining the output of viewpoint-specific detectors with a linear SVM (bottom).

%	Right	Right-Back	Back	Left-Back	Left	Left-Front	Front	Right-Front	**Avg.**
Max	53.7	35.5	45.7	22.6	37.9	8.6	40.0	8.3	**31.1**
SVM	72.6	12.7	48.6	12.3	55.7	44.5	70.4	16.2	**42.2**
SVM-adj	71.4	22.3	29.5	18.0	84.7	18.1	50.7	29.2	**35.4**

Table 1. Viewpoint estimation on the "MultiviewPeople" dataset. The task is to classify one of 8 viewpoints (chance level 12.5%).

Viewpoint estimation. Next we aim to estimate the person's viewpoints, since such viewpoint estimates allow to significantly reduce the ambiguity in 3D pose. To that end, we rely on the bank of viewpoint-specific detectors from above, and train 8 viewpoint classifiers, linear SVMs, on the detector outputs of the validation set. We consider two training and evaluation strategies: (*SVM*) Only training examples from one of the viewpoints are used as positive examples, the remainder as negative ones; and (*SVM-adj*), where we group viewpoints into triplets of adjacent ones and train separate classifiers for each such triplet. As a baseline approach (*Max*), we estimate the viewpoint by taking the maximum over the outputs of the 8 viewpoint-specific detectors. Results are shown in Tab. 1. *SVM* improves over the baseline in case when we require exact recognition of viewpoint by approx. 11%, but *SVM-adj* also performs well. In addition, when we also consider the two adjacent viewpoints as being correct, *SVM* obtains an average performance of 70.0% and *SVM-adj* of 76.2%. This shows that *SVM-adj* more gracefully degrades across viewpoints, which is why we adopt it in the remainder.

Since the scores of the viewpoint classifiers are not directly comparable with each other, we calibrate them by computing the posterior of the correct label given the classifier score, which maps the scores to the unit interval. The posterior is computed via Bayes' rule from the distributions of classifier scores on the positive and negative examples. We assume these distributions to be Gaussian, and estimate their parameters from classifier scores on the validation set.

Fig. 3 shows calibrated outputs of all 8 classifiers computed for a sequence of 40 frames in which the person first appears from the "right" and then from the "right-back" viewpoint. The correct viewpoint is the most probable for most of the sequence, and failures in estimation often correspond to adjacent viewpoints.

3. 2D Tracking and Viewpoint Estimation

As discussed in the introduction, our goal is to accumulate all available 2D image evidence prior to the third 3D tracking stage in order to reduce the ambiguity of 2D-to-3D lifting as much as possible. While the person detector described in the previous section is capable of estimating 2D positions of body parts and viewpoints of people from single frames, the second stage (described here) aims to improve these estimates by 2D-tracking-by-detection [2, 31].

To exploit temporal coherency already in 2D, we extract short tracklets of people. This, on the one hand, improves the robustness of estimates for 2D positions, scale and viewpoint of each person, since they are jointly estimated over an entire tracklet. Improved body localization in turn aids 2D pose estimation. On the other hand, it also allows to perform *early data association*. This is important for sequences with multiple people, where we can associate "anonymous" single frame hypotheses with the track of a specific person.

Tracklet extraction. From the first stage of our approach we obtain a set of N_m potentially overlapping bounding box hypotheses $\mathcal{H}_m = [\mathbf{h}_{m1}, \ldots, \mathbf{h}_{mN_m}]$ for each frame m of the sequence, where each hypothesis $\mathbf{h}_{mi} = \{h^x_{mi}, h^y_{mi}, h^s_{mi}\}$ corresponds to a bounding box at particular image position and scale. In order to obtain a set of tracklets, we follow the HMM-based tracking procedure introduced in [2][1]. To that end we treat the person hypotheses in each frame as states and find state subsequences that

[1] A detailed description is given in Sec. 3.3 of [2].

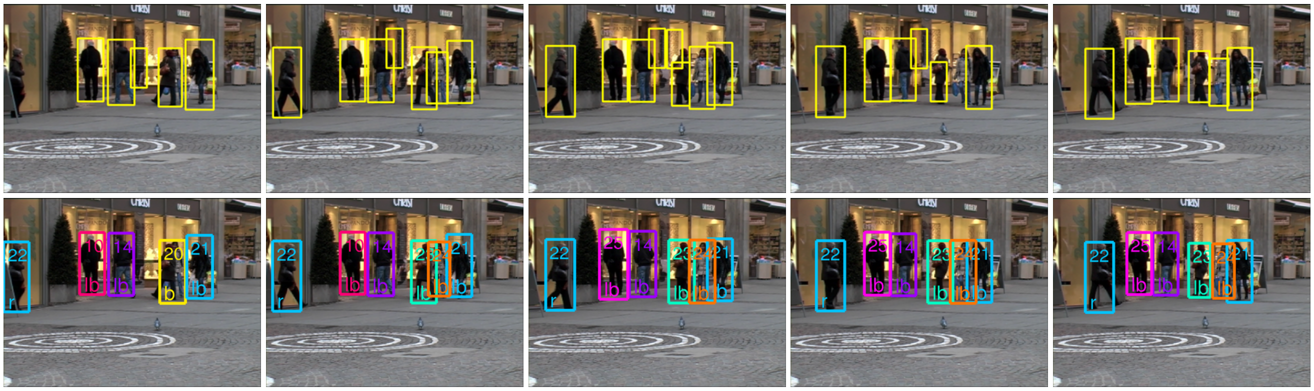


Figure 5. People detection based on single frames (top) and tracklets found by our 2D tracking algorithm; Different tracklets are identified by color and the estimated viewpoints are indicated with two letters (bottom). Note that several false positives in the top row are filtered out and additional – often partially occluded – detections are filled in (e.g., on the left side of the leftmost image).

are consistent in position, scale and appearance by iteratively applying Viterbi decoding. The emission probabilities for each state are derived from the detection score. The transition probabilities between states $\mathbf{h}_{mi}$ and $\mathbf{h}_{m-1,j}$ are modeled using first-order Gaussian dynamics and appearance compatibility:

$$\begin{aligned} p_{trans}(\mathbf{h}_{mi},\mathbf{h}_{m-1,j}) \quad = \quad & \mathcal{N}(\mathbf{h}_{mi}|\mathbf{h}_{m-1,j},\Sigma_{pos}) \cdot \\ & \mathcal{N}(d_{app}(\mathbf{h}_{mi},\mathbf{h}_{m-1,j})|0,\sigma^2_{app}). \end{aligned}$$

where $\Sigma_{pos} = \text{diag}(\sigma_x^2, \sigma_y^2, \sigma_s^2)$, and $d_{app}(\mathbf{h}_{mi},\mathbf{h}_{m-1,j})$ is the Euclidean distance between RGB color histograms computed for the bounding rectangle of each hypothesis. We set $\sigma_x = \sigma_y = 5$, $\sigma_s = 0.1$ and $\sigma_{app} = 0.05$.

Viewpoint tracking. Finally, for each of the tracklets we estimate the viewpoint sequence $\omega_{1:N} = (\omega_1, \ldots, \omega_N)$, again using a simple HMM and Viterbi decoding. We consider the 8 discrete viewpoints as states, the viewpoint classifiers described in Sec. 2 as unary evidence, and Gaussian transition probabilities that enforce similar subsequent viewpoints to reflect that people tend to turn slowly.

Evaluation. Fig. 5 shows an example of a short subsequence in which we compare the detection results of the single-frame 2D detector with the extracted tracklets. Note how tracking helps to remove the spurious false positive detections in the background, and corrects failures in scale estimation, which would otherwise hinder correct 2D-to-3D lifting. In Fig. 3 we visualize the single frame prediction scores for each viewpoint for the tracklet corresponding to the person with index 22 on Fig. 5. Note that while viewpoint estimation from single frames is reasonably robust, it can still fail at times (the correct viewpoint is "right" for frames 4 to 30 and "right-back" for frames 31 to 40). The tracklet-based viewpoint estimation, in contrast, yields the correct viewpoint for the entire 40 frame sequence. Finally, as we demonstrate in Fig. 5, the tracklets also provide data association even in case of realistic sequences with frequent full and partial occlusions.

4. 3D Pose Estimation

To estimate and track poses in 3D, we take the 2D tracklets extracted in the previous stage and lift the 2D pose estimated for each frame into 3D (c.f. [25]), which is done with the help of a set of 3D exemplars [19, 22]. Projections of the exemplars are first evaluated under the 2D body part posteriors, and the exemplar with the most probable projection is chosen as an initial 3D pose. This initial pose is propagated to all frames of the tracklet using the known temporal ordering on the exemplar set. Note that this yields multiple initializations for 3D pose sequences, one for each frame of the tracklet. This 2D-to-3D lifting procedure is robust, because it is based on reliable 2D pose posteriors, and detections and viewpoint estimates from the 2D tracklets. Starting from these initial pose sequences, the actual pose estimation and tracking is done in a Bayesian framework by maximizing the posterior defined in Eq. (1), for which they serve as powerful initializations. The 3D pose is parametrized as $Q_m = \{\mathbf{q}_m, \phi_m, \mathbf{h}_m\}$, where $\mathbf{q}_m$ denotes the parameters of body joints, ϕ_m the rotation of the body in world coordinates, and $\mathbf{h}_m = \{h_m^x, h_m^y, h_m^{scale}\}$ the position and scale of the person projected to the image. The 3D pose is represented using a kinematic tree with $P = 10$ flexible joints, in which each joint has 2 degrees of freedom. A configuration example is shown in Fig. 6(a).

The evidence at frame m is given by the $\mathcal{E}_m = \{D_m, \omega_m\}$, and consists of the single frame image evidence D_m and the 2D viewpoint estimate ω_m obtained from the entire tracklet. Assuming conditional independence of 2D viewpoint and image evidence given the 3D pose, the single frame likelihood in Eq. 2 factorizes as

$$p(\mathcal{E}_m|Q_m) = p(\omega_m|Q_m)p(D_m|Q_m). \qquad (6)$$

Based on the estimated 2D viewpoint ω_m, we model the viewpoint likelihood of the 3D viewpoint ϕ_m as Gaussian centered at the rotation component of ϕ_m along the y-axis:

$$p(\omega_m|Q_m) = \mathcal{N}(\omega_m|\text{proj}_y(\phi_m), \sigma_\omega^2). \qquad (7)$$

We define the likelihood of the 3D pose $p(D_m|Q_m)$ with the help of the part posteriors given by the 2D body model

$$p(D_m|Q_m) = \prod_{n=1}^{N} p(\text{proj}_n(Q_m)|D_m), \tag{8}$$

where $\text{proj}_n(Q_m)$ denotes the projection of the n-th 3D body part into the image. While such a 3D likelihood is typically defined as the product of individual part likelihoods similarly to Eq. 5, this leads to highly multimodal posteriors and difficult inference. By relying on 2D part posteriors instead, the 3D model is focused on hypotheses for which there is sufficient 2D image evidence from previous stages.

To avoid expensive 3D likelihood computations, we represent each 2D part posterior using a non-parametric representation. In particular, for each body part n in frame m we find the J locations with the highest posterior probability $E_{mn} = \{(\mathbf{l}_{mn}^j, w_{mn}^j), j = 1, \ldots, J\}$, where $\mathbf{l}_{mn}^j \in \mathbb{R}^4$ corresponds to the 2D location (image position and orientation) and w_{mn}^j to the posterior density at this location. Given that, we approximate the 2D part posterior as a kernel density estimate with Gaussian kernel κ:

$$p(\text{proj}_n(Q_m)|D_m) \approx \sum_j w_{mn}^j \kappa(\mathbf{l}_{mn}^j, \text{proj}_n(Q_m)). \tag{9}$$

Dynamical model. In our approach we represent the temporal prior in Eq. 1 as the product of two terms:

$$p(Q_{1:M}) = p(\mathbf{q}_{1:M})p(\mathbf{h}_{1:M}), \tag{10}$$

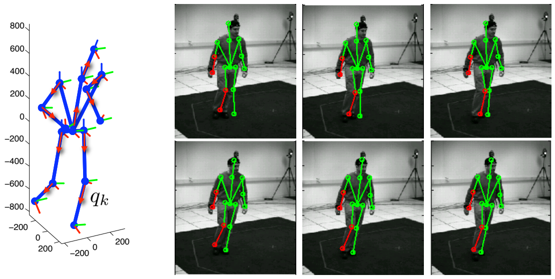
which correspond to priors on the parameters of 3D pose as well as image position and scale. The prior on the person's position and scale $p(h_{1:M})$ is taken to be a broad Gaussian and models smooth changes of both the scale of the person and its position in the image.

We model the prior over the parameters of the 3D pose $\mathbf{q}_{1:M}$ with a hierarchical Gaussian process latent variable model (hGPLVM) [17]. We denote the M dimensional vector of the values of i-th pose parameter across all frames as $\mathbf{q}_{1:M,i}$. In the hGPLVM each dimension of the original high-dimensional pose is modelled as an independent Gaussian process defined over a shared low dimensional latent space $\mathbf{Z}_{1:M}$:

$$p(\mathbf{q}_{1:M}|\mathbf{Z}_{1:M}) = \prod_{i=1}^{P} \mathcal{N}(\mathbf{q}_{1:M,i}|0, \mathbf{K}_z), \tag{11}$$

where P is the number of parameters in our pose representation, $\mathbf{K}_z$ is a covariance matrix of the elements of the shared latent space $\mathbf{Z}_{1:M}$ defined by the output of the covariance function $\text{k}(\mathbf{z}_i, \mathbf{z}_j)$, which in our case is taken to be squared exponential.

The values of the shared latent space $\mathbf{Z}_{1:M}$ are themselves treated as the outputs of Gaussian processes with a one dimensional input, the time $T_{1:M}$. Our implementation uses a $d_l = 2$ dimensional shared latent space. Such a hierarchy of Gaussian processes allows to effectively model both correlations between different dimensions of the original input space and their dynamics.



(a)

(b)

Figure 6. (a): Representation of the 3D pose in our model (parametrized joints are marked with arrows). (b): Initial pose sequence after 2D-to-3D lifting (top) and pose sequence after optimization of the 3D pose posterior (bottom).

MAP estimation. The hGPLVM prior requires two sets of auxiliary variables $\mathbf{Z}_{1:M}$ and $T_{1:M}$, which need to be dealt with during maximum a-posteriori estimation. Our strategy is to optimize $\mathbf{Z}$ only and keep the values of T fixed. This is possible since the values of T roughly correspond to the person's state within a walking cycle, which can be reliably estimated using the 2D tracklets. The full posterior over 3D pose parameters being maximized is given by:

$$\begin{aligned} p(Q_{1:M}, \mathbf{Z}_{1:M}|\mathcal{E}_{1:M}, T_{1:M}) &\propto p(\mathcal{E}_{1:M}|Q_{1:M}) \cdot \\ & p(\mathbf{q}_{1:M}|\mathbf{Z}_{1:M})p(\mathbf{Z}_{1:M}|T_{1:M})p(h_{1:M}). \end{aligned} \tag{12}$$

We optimize the posterior using scaled conjugate gradients and initializing the optimization using the lifted 3D poses.

4.1. 3D pose estimation for longer sequences

The MAP estimation approach just described is only tractable if the number of frames M is sufficiently small. In order to estimate 3D poses in longer sequences we apply a strategy similar to the one used in [2]: First we estimate 3D poses in short ($M = 10$) overlapping subsequences of the longer sequence. Since for each subsequence we initialize and locally optimize the posterior multiple times, this leaves us with a large pool of 3D pose hypotheses for each of the frames, from which we find the optimal sequence using a hidden Markov model and Viterbi decoding. To that end we treat the 3D pose hypotheses in each frame as discrete states with emission probabilities given by Eq. 6 and define transition probabilities between states using the hGPLVM.

5. Experiments

We evaluate our model in two diverse scenarios. First, we show that our approach improves the state-of-the-art in monocular human pose estimation on the standard "HumanEva II" benchmark, for which ground truth poses are available. Additionally, we evaluate our approach on two cluttered and complex street sequences with multiple people including partial and full occlusions.

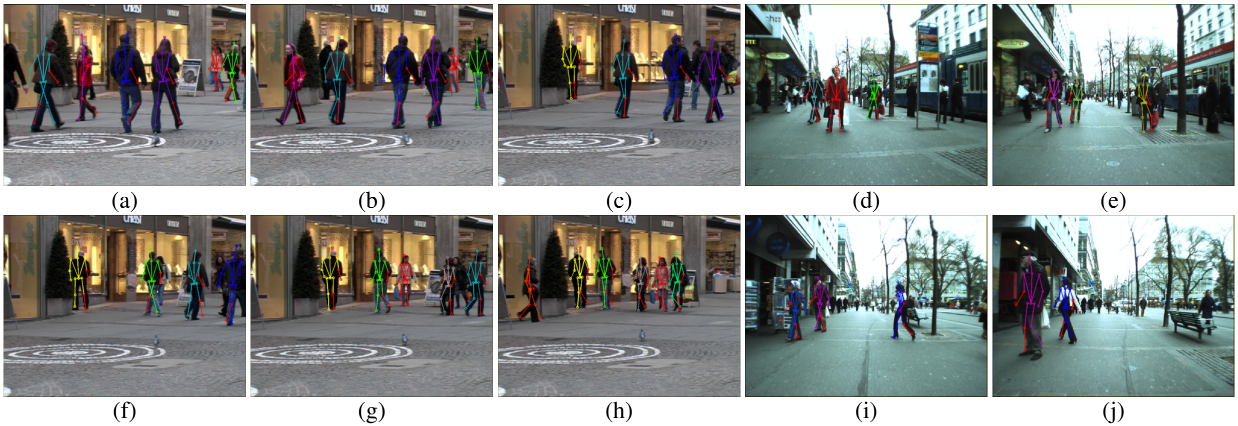


Figure 7. 3D pose estimation on the "TUD Stadtmitte" dataset (left) and on a sequence from a moving camera [14] (right).

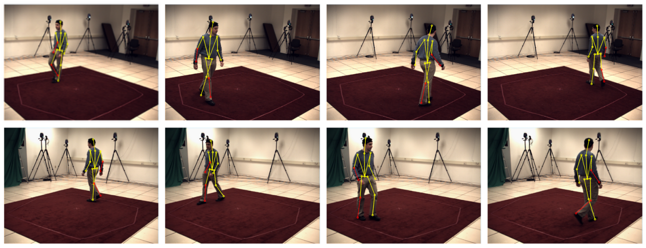


Figure 8. 3D pose estimation examples on HumanEva II for "Subject S2/Camera C1" (top) and "Subject S2/Camera C2" (bottom).

5.1. Evaluation on the "HumanEva II" dataset

In order to quantitatively evaluate the performance of our 3D pose estimation method we use the "HumanEva II" dataset [23], which provides synchronized images and motion capture data and is a standard evaluation benchmark for 2D and 3D human pose estimation. On this dataset we compare to [21] as they obtain the best published results in a setting comparable to ours: They estimate poses in monocular image sequences without background subtraction, but rely on both appearance and temporal information.

For this experiment we train viewpoint specific models on the images of subjects "S1", "S2", "S3" from the "HumanEva" dataset. We found that adding more training data improves the performance of part detectors, especially for the lower and upper arm body parts. Therefore, we extended the training data with the images from the "People" [20] and "Buffy" [10] datasets. The set of exemplars for 2D-to-3D lifting and the hGPLVM used to model the temporal dynamics on the pose sequences were obtained using training data for subject "S3" of the "HumanEva" dataset. As we show, despite the limited training data, this prior enables pose estimation on the "HumanEva II" dataset as well as in realistic street scenes.

The authors of [21] report pose estimation results for the first 350 frames of the sequence containing subject "S2", independently estimating poses for views obtained from cameras "C1" and "C2". Tab. 2 shows the mean error in the estimation of 2D and 3D joint locations, obtained using the official online evaluation tool. For both sequences our results improve substantially over those reported by [21]. The improvement is especially large for the sequence taken with camera "C2", on which we obtain an average error of 10.72 pixels, compared to 14.18 pixels. We have also evaluated the 3D pose estimation performance of our approach, obtaining a mean error of 107 and 101 millimeters for cameras "C1" and "C2". Fig. 8 shows several examples of estimated poses obtained by our method on both sequences, visualizing every 100-th frame. We attribute the better localization accuracy of our method to the continuous optimization of 3D pose given the body part positions rather than selecting one from a discrete set of exemplars [21].

Subj./Cam.	2D Mean (Std) [21]	2D Mean (Std)	3D Mean (Std)
S2/C1	12.98(3.5)	**10.49** (**2.70**)	107(15)
S2/C2	14.18(4.38)	**10.72** (**2.44**)	101(19)

Table 2. Quantitative evaluation on the HumanEva II dataset (frames 1-350). We report mean error and standard deviation of the relative 2D and 3D joint positions. 2D results are in pixels and 3D results are in millimeters.

5.2. 3D pose estimation in street scenes

To evaluate our approach for a realistic street setting, we introduce the novel "TUD Stadtmitte"' dataset containing 200 consecutive frames taken in a typical pedestrian area. Over the 200 frames of this sequence our 2D tracking algorithm obtained 25 2D people-tracks, none of which contained false positive detections. Fig. 7 (left) shows example images evenly spaced throughout the sequence. For every track we estimate the viewpoint of the person using exclusively our viewpoint classification algorithm. We could easily integrate the direction of movement of the person into the estimate, but this would limit the applicability of our method to the setting with static cameras. Note that we are able to estimate 3D poses correctly over a diverse range of viewpoints including those with significant depth ambiguity and difficult imaging conditions. Note, for example, the people on the right side of image Fig. 7(a) and the people in the middle of the image in Fig. 7(h).

Unfortunately we cannot report quantitative results on the 3D pose recovery for this sequence, as obtaining ground truth is very difficult for such realistic image sequences. However, the results are qualitatively close to those demonstrated by our method on the "HumanEva II" dataset, suggesting that the obtainable quantitative results would be comparable. Although our motion prior was trained on the "HumanEva" dataset, it generalized well to the street setting. Interestingly, we are also able to correctly estimate the pose of the person standing still, as is shown in Fig. 7(c,f,g). Looking at typical failure cases, several incorrectly estimated poses are due to incorrect scale estimation as in Fig. 7(a), partial occlusion Fig. 7(b,f), or failure in viewpoint estimation (e.g., the rightmost person in Fig. 7(h)).

We also evaluated our approach on a sequence recorded by a moving camera previously used in [14]. Due to the high amount of background clutter, low frame-rate and many people in near frontal views, this sequence presents significant challenges for 3D pose estimation. Several examples of estimated 3D poses are shown in Fig. 7(right). Note that even under such challenging conditions our approach can track and estimate poses of people over a large number of frames, e.g., the rightmost person in Fig. 7(i,j). Also note that tracking and viewpoint estimation produced correct results even in the presence of strong background clutter, e.g., for the rightmost person in Fig. 7(d,e).

6. Conclusions

In this paper we presented a novel approach to monocular 3D human pose estimation and tracking, which is able to recover poses of people in realistic street conditions. The approach leverages recent advances in reliable 2D pose estimation from monocular images, tracking-by-detection, and powerful modeling of 3D dynamics based on hierarchical Gaussian process latent variable models. This allows to first accumulate the available 2D image evidence from which later 3D poses can be reliably recovered and tracked. The approach has been evaluated quantitatively on the "HumanEva II" benchmark and improves the state-of-the-art in this setting. We also showed excellent results on a challenging street sequence underlining the applicability of the approach for 3D pose estimation and tracking of multiple people in cluttered scenes using a monocular, potentially moving camera.

Acknowledgements: Mykhaylo Andriluka gratefully acknowledges a scholarship from DFG GRK 1362.

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] M. Andriluka, S. Roth, and B. Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *CVPR-09*.

[4] A. Balan, L. Sigal, M. Black, J. Davis, and H. Haussecker. Detailed human shape and pose from images. In *CVPR-07*.

[5] S. Belongie, J. Malik, and J. Puzicha. Shape context: A new descriptor for shape matching and object recognition. In *NIPS*00*.

[6] J. Deutscher and I. Reid. Articulated body motion capture by stochastic search. *IJCV*, 61:185–205, Feb. 2005.

[7] M. Eichner and V. Ferrari. Better appearance models for pictorial structures. In *BMVC-09*.

[8] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *IJCV*, 61:55–79, Jan. 2005.

[9] P. F. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *CVPR-08*.

[10] V. Ferrari, M. Marin, and A. Zisserman. Progressive search space reduction for human pose estimation. In *CVPR-08*.

[11] A. Fossati, M. Dimitrijevic, V. Lepetit, and P. Fua. Bridging the gap between detection and tracking for 3D monocular video-based motion capture. In *CVPR-07*.

[12] Y. Freund and R. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *J. of Comp. and Sys. Sc.*, 55(1):119–139, 1997.

[13] J. Gall, B. Rosenhahn, T. Brox, and H.-P. Seidel. Optimization and filtering for human motion capture: A multi-layer framework. *IJCV*, 87(1–2), Mar. 2010.

[14] S. Gammeter, A. Ess, T. Jaeggli, K. Schindler, B. Leibe, and L. Gool. Articulated multi-body tracking under egomotion. In *ECCV-08*.

[15] N. Hasler, B. Rosenhahn, T. Thormaehlen, M. Wand, and H.-P. Seidel. Markerless motion capture with unsynchronized moving cameras. In *CVPR-09*.

[16] C. Ionescu, L. Bo, and C. Sminchisescu. Structural SVM for visual localization and continuous state estimation. In *ICCV-09*.

[17] N. D. Lawrence and A. J. Moore. Hierarchical Gaussian process latent variable models. In *ICML-07*.

[18] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *PAMI*, 27(10):1615–1630, 2005.

[19] G. Mori and J. Malik. Recovering 3D human body configurations using shape contexts. *PAMI*, 28(7):1052–1062, 2006.

[20] D. Ramanan. Learning to parse images of articulated objects. In *NIPS*06*.

[21] G. Rogez, J. Rihan, S. Ramalingam, C. Orrite, and P. H. Torr. Randomized trees for human pose detection. In *CVPR-08*.

[22] G. Shakhnarovich, P. A. Viola, and T. Darrell. Fast pose estimation with parameter-sensitive hashing. In *ICCV-03*.

[23] L. Sigal and M. Black. HumanEva: Synchronized video and motion capture dataset for evaluation of articulated human motion. Technical report, Brown University, 2006.

[24] L. Sigal and M. J. Black. Measure locally, reason globally: Occlusion-sensitive articulated pose estimation. In *CVPR-06*.

[25] L. Sigal and M. J. Black. Predicting 3D people from 2D pictures. In *AMDO 2006*.

[26] Z. Tu, X. Chen, A. L. Yuille, and S.-C. Zhu. Image parsing: Unifying segmentation, detection, and recognition. *IJCV*, 63(2):113–140, 2005.

[27] R. Urtasun and T. Darrell. Local probabilistic regression for activity-independent human pose inference. In *ICCV-09*.

[28] R. Urtasun, D. J. Fleet, and P. Fua. 3D people tracking with Gaussian process dynamical models. In *CVPR-06*.

[29] M. Vondrak, L. Sigal, and O. C. Jenkins. Physical simulation for probabilistic motion tracking. In *CVPR-08*.

[30] C. Wojek, S. Walk, and B. Schiele. Multi-cue onboard pedestrian detection. In *CVPR-09*.

[31] B. Wu and R. Nevatia. Detection and tracking of multiple, partially occluded humans by Bayesian combination of edgelet based part detectors. *IJCV*, 75:247–266, Nov. 2007.