People-Tracking-by-Detection and People-Detection-by-Tracking

Mykhaylo Andriluka Stefan Roth Bernt Schiele
Computer Science Department
Technische Universität Darmstadt
{andriluka, sroth, schiele}@cs.tu-darmstadt.de

Abstract

Both detection and tracking people are challenging problems, especially in complex real world scenes that commonly involve multiple people, complicated occlusions, and cluttered or even moving backgrounds. People detectors have been shown to be able to locate pedestrians even in complex street scenes, but false positives have remained frequent. The identification of particular individuals has remained challenging as well. On the other hand, tracking methods are able to find a particular individual in image sequences, but are severely challenged by real-world scenarios such as crowded street scenes. In this paper, we combine the advantages of both detection and tracking in a single framework. The approximate articulation of each person is detected in every frame based on local features that model the appearance of individual body parts. Prior knowledge on possible articulations and temporal coherency within a walking cycle are modeled using a hierarchical Gaussian process latent variable model (hGPLVM). We show how the combination of these results improves hypotheses for position and articulation of each person in several subsequent frames. We present experimental results that demonstrate how this allows to detect and track multiple people in cluttered scenes with reoccurring occlusions.

1. Introduction

This paper addresses the challenging problem of detection and tracking of multiple people in cluttered scenes using a monocular, potentially moving camera. This is an important problem with a wide range of applications such as video indexing or surveillance of airports and train stations. Probably the most fundamental difficulty in detection and tracking many people in cluttered scenes is that many people will be partially and also fully occluded for longer periods of times. Consequently, both the detection of people in individual frames as well as the data-association between people detections in different frames are highly challenging and ambiguous. To address this, we exploit temporal coherency, extract people-tracklets from a small number of consecutive frames and from those tracklets build models of the individual people. As any single person might be detectable only for a small number of frames the extraction of people-tracklets has to be highly robust. At the same time the extracted model of the individual has to be discriminative enough in order to enable tracking and data-association across long periods of partial and full occlusions.

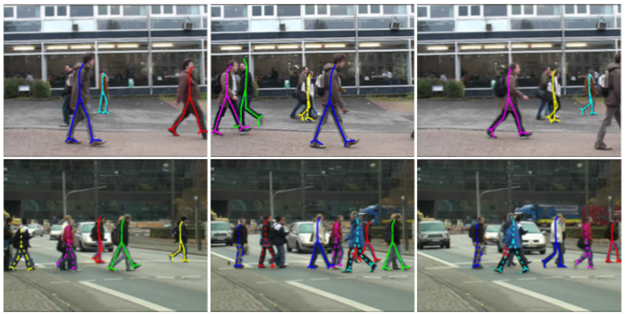


Figure 1. Examples of detection and tracking of *specific* persons in image sequences of crowded street scenes.

To achieve reliable extraction of people-tracklets as well as data-association across long periods of occlusion, the proposed approach combines recent advances in people detection with the power of dynamical models for tracking. Rather than to simply determine the position and scale of a person as is common for state-of-the-art people detectors [3, 16], we also extract the position and articulation of the limbs. This allows us to use a more powerful dynamical model that extends people detection to the problem of reliably extracting people-tracklets – people detections consistent over a small number of frames. In particular, we use a hierarchical Gaussian process latent variable model (hGPLVM) [14] to model the dynamics of the individual limbs. As we will show in the experiments this enables us to detect people more reliably than it would be possible from single frames alone. 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 people-tracks. These people-tracks identify individuals over longer sequences of consecutive frames when that is appropriate, such as between major occlusion events. Tracking people over even longer periods of time is then achieved by associating people-tracks across potentially long periods of occlusion using both the dynamical model and an extracted appearance model, which allows identifying specific individuals throughout the sequence.

The first contribution of this paper is the extension of a state-of-the-art people detector [16] with a limb-based structure model. Sec. 2 shows that this novel detector outperforms two state-of-the-art detectors on a challenging dataset. The novel detector has two important properties that make it particularly suited for the detection and tracking of multiple people in crowded scenes: First, it allows to detect people in the presence of significant partial occlusions, and second, the output of the detector includes the positions of individual limbs, which are used as input for the dynamical model at the next stage. The second contribution of the paper is to integrate the people detection model using a dynamical limb-model to enable reliable detection of people-tracklets over small number of consecutive frames, which further improves the detection performance. To our knowledge, this is also the first application of the hGPLVM dynamical model to a complex vision problem. The extracted people-tracklets are then used to generate a detailed appearance model of each person on the fly. The third contribution is to link short people-tracklets to longer tracks of the various individuals in the scene. In this, we take advantage of the articulated people detector, which allows us to do filtering using a HMM model with a simple discrete state space. This is in contrast to typical tracking approaches that need to perform stochastic search in high-dimensional, continuous spaces [5], which is well known to suffer from many problems [4, 22]. Note that while a precise recovery of the full articulation is not the main focus of this paper, we can still quite accurately recover the articulation even in complex scenes. The final contribution of the paper is to associate people-tracks in scenes with multiple people and over periods of long occlusions. In particular, the learned appearance model allows us to identify and track individuals even through complex occlusions without requiring any manual initialization or manual intervention at *any* stage of the process.

1.1. Related work

Tracking by detection has been a focus of recent work [18, 8, 27, 1]. This research has been facilitated by the impressive advances in people detection methods [24, 3, 16]. Approaches most related to what is proposed in this paper include the work by Leibe *et al.* [15], who have extended [16] to enable detection and trajectory estimation in complex traffic scenes. This approach, however, has not been shown be able to handle complex and long occlusions as we focus on here. Ramanan *et al.* [19] propose a two-stage approach that first builds a model of the appearance of individual people, and then tracks them by detecting those models in each frame. Their approach uses only very simple limb detectors based on finding parallel lines of contrast in the image. Here, we use more refined limb detectors that are learned, yet generic enough to detect peoples' limbs in a wide variety of situations. Wu and Nevatia [27] propose an approach for detecting and tracking partially occluded people using an assembly of body parts. All of these approaches explicitly or implicitly assume that humans are either only partially occluded or fully occluded only for short periods of time. In this sense this paper pushes the state-of-art by addressing the important and frequent problem of significant and long-term occlusions for crowded scenes with many people.

A review of the literature on people tracking is well beyond the scope of this paper, and hence we will only mention a few examples of related work here. Many approaches, even to this date, are based on silhouettes (*e.g.* [5]) and perform tracking using stochastic search in high-dimensional spaces. While using silhouettes may be appropriate for tracking a single person, silhouette extraction becomes unreliable because of complex backgrounds, occlusions, and moving cameras. Moreover, stochastic search in these high-dimensional spaces is notoriously difficult. To work around these problems, a number of recent tracking approaches turned to feature-based detectors for matching tracking hypotheses, discriminative components, strong dynamical models, or alternative methods for exploring the search space [4, 23, 7, 22]. Fossati *et al.* [7], for example, perform tracking aided by detection, even in 3D, but need to find a ground plane and only track single individuals without substantial occlusions. Sminchisescu *et al.* [22] combine discriminative prediction of the body state from densely sampled features with a dynamical model. Their method focuses on the accurate recovery of the articulation of the single person, whereas we focus on the robust detection and tracking of multiple people in scenes with complex, long-term occlusions. Sigal and Black [21], for example, integrate occlusion reasoning into a 2D articulated tracking model. Their model only deals with self-occlusions, however. This paper instead focuses on occlusions between different people as well as on people being occluded by the environment. In that, this paper is also related to multiple-people blob-tracking methods, such as [11], but we do not need to assume a static camera and allow for low viewpoints (also in contrast to [15]) from which people can fully occlude each other.

2. Pedestrian Detector

Before introducing temporal constraints into the detection process, we first propose a novel part-based object detection model that is capable of detecting pedestrians in sin-

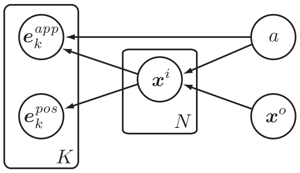


Figure 2. Graphical model structure describing the relation between articulation, parts, and features.

gle images of real-world scenes. The model is inspired by the pictorial structures model proposed by [6, 10], but uses more powerful part representations and detection, and as we will show outperforms recent pedestrian detectors [3, 20].

2.1. Part-based model for pedestrian detection

Following the general pictorial structures idea, an object is represented as a joint configuration of its parts. In such a model the problem of locating an object from a specific class in a test image is formulated as search for the modes of the posterior probability distribution $p(L|E)$ of the object configuration L given the image evidence E and (implicit) class-dependent model parameters θ.

In our model, the configuration is described as $L = \{\mathbf{x}^o, \mathbf{x}^1, \dots, \mathbf{x}^N\}$, where $\mathbf{x}^o$ is the position of the object center and its scale, and $\mathbf{x}^i$ is the position and scale of part i. The image evidence, which here is defined as a set of local features observed in the test image, will be denoted as $E = \{\mathbf{e}_k^{app}, \mathbf{e}_k^{pos} | k = 1, \dots, K\}$, where $\mathbf{e}_k^{app}$ is an appearance descriptor, and $\mathbf{e}_k^{pos}$ is the position and scale of the local image feature with index k. We will denote the combination of position, scale, and appearance of a local feature as $\mathbf{e}_k = (\mathbf{e}_k^{app}, \mathbf{e}_k^{pos})$.

An important component of the pictorial structures model is an implicit model of a-priori knowledge about possible object configurations, which must be expressive enough to capture all important dependencies between parts. Part positions are mutually dependent in general, which can make inference difficult. But for particular object categories, such as walking people, we can introduce auxiliary state variables that represent the *articulation state* or an *aspect* of the object, such as different phases in the walking cycle of a person [12], and make the parts conditionally independent. If the articulation state is observed, the model becomes a star model (or tree model in general) and efficient algorithms based on dynamic programming [6] can be used for inference. If we are not interested in knowing the articulation state, but only the object and limb positions, then articulation state a can be marginalized out:

$$p(L|E) = \sum_a p(L|a, E) p(a). \qquad (1)$$

From decomposing $p(L|a, E) \propto p(E|L, a) p(L|a)$, assuming that the configuration likelihood can be approximated with product of individual part likelihoods [6] $p(E|L, a) \approx \prod_i p(E|\mathbf{x}^i, a)$, and assuming uniform $p(\mathbf{x}^i|a)$, it follows that

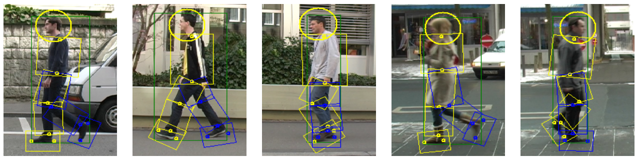


Figure 3. Examples of images from our training set.

$$p(L|a, E) \approx p(\mathbf{x}^o) \prod_i p(\mathbf{x}^i|a, E) p(\mathbf{x}^i|\mathbf{x}^o, a). \qquad (2)$$

If we assume that a particular image feature $\mathbf{e}_k$ belongs to part i of an object instance in the image with probability α, then it holds that

$$p(\mathbf{x}^i|a, E) = c_0 + c_1 \sum_{\mathbf{e}_k} p(\mathbf{x}^i|a, \mathbf{e}_k) + O(\alpha^2), \qquad (3)$$

where c_0 and c_1 depend only on the image features E [26]. If α is sufficiently small, which is true for street scenes in which a particular person usually represents only a small portion of the image, we obtain

$$p(L|a, E) \approx \prod_i p(\mathbf{x}^i|\mathbf{x}^o, a) \left[\beta + \sum_{\mathbf{e}_k} p(\mathbf{x}^i|a, \mathbf{e}_k) \right], \qquad (4)$$

where β can be seen as a regularizer for the evidence obtained from the individual image features, and we have additionally assumed uniform $p(\mathbf{x}^o)$.

As is common in models based on local feature representations, we introduce an object-specific *codebook* denoted as $\mathcal{C} = \{\mathbf{c}_j | j = 1, \dots, J\}$. The part posterior with respect to a single image feature is computed by marginalization over the codebook entries:

$$p(\mathbf{x}^i|a, \mathbf{e}_k) = \sum_{\mathbf{c}_j} p(\mathbf{x}^i|a, \mathbf{c}_j, \mathbf{e}_k^{pos}) p(\mathbf{c}_j|\mathbf{e}_k^{app}). \qquad (5)$$

$p(\mathbf{c}_j|\mathbf{e}_k^{app})$ is discrete distribution over codebooks based on a Gaussian similarity measure, and $p(\mathbf{x}^i|a, \mathbf{c}_j, \mathbf{e}_k^{pos})$ is learned from training data (see below). The structure of the dependencies between the variables in the model is shown in Fig. 2.

2.2. Model training

In all presented experiments we use shape context feature descriptors [2] and the Hessian-Laplace interest point operator [17] as detector. The object-specific codebook is constructed by clustering local features extracted from the set of training images.

For each codebook cluster $\mathbf{c}_j$ we also compute its *occurrence distribution*, which corresponds to a set of jointly

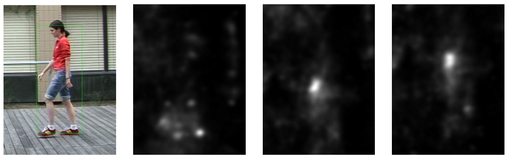


Figure 4. Person hypothesis with corresponding probabilities of foot, upper leg, and torso.

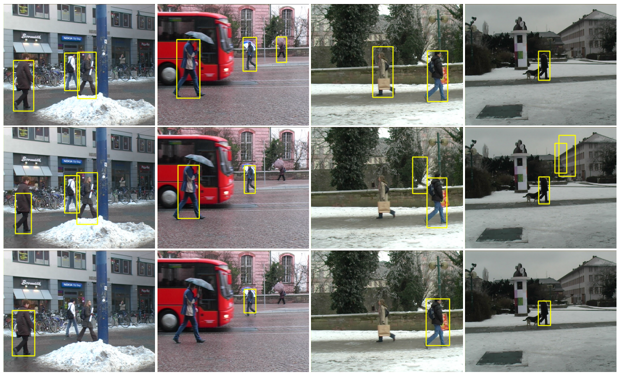


Figure 5. Example detections at equal error rate of our detector (top), 4D-ISM (middle) and HOG (bottom) on the "TUD-Pedestrians" dataset.

observed relative position and scale of the cluster with respect to the part centers computed for each occurrence of the cluster in the training set. This allows us to compute $p(\mathbf{x}^i|a, \mathbf{c}_j, \mathbf{e}_k^{pos})$.

In order to compute the occurrence distributions we manually annotated each person along with its parts in all training images. Fig. 3 shows several images from our training set. From the same manual annotation we also learn a Gaussian distribution of the position of each part relative to the object center, $p(\mathbf{x}^i|\mathbf{x}^o, a)$. While the position components are learned, the scale component is taken to be relatively broad and chosen empirically. This appears to be sufficient for pedestrians, particularly since we are not differentiating between left and right legs at the detection stage.

2.3. Inference and Results

In the first step of inference we accumulate $p(\mathbf{x}^i|a, \mathbf{e}_k)$ in a 3 dimensional array of discretized image positions and scales. After that Eq. (4) can be maximized efficiently using the generalized distance transform [6]. This is possible since part dependencies have a tree (star-) structure, appearance components are computed separately for each part, and $p(\mathbf{x}^i|\mathbf{x}^o, a)$ is Gaussian. Fig. 4 visualizes $\sum_{\mathbf{e}_k} p(\mathbf{x}^i|a, \mathbf{e}_k)$ in the region of a person hypothesis for different limbs.

In the following we evaluate the novel detector on a challenging dataset of 250 images of street scenes containing 311 side-view pedestrians with significant variation in clothing and articulation, which we denote as "TUD-Pedestrians" [1]. Fig. 7(a) shows the comparison of our detector with two state of the art detectors. Using the same training set as [20] our detector outperforms the 4D-ISM approach [20] as well as the HOG-detector [3]. Increasing the size of the training set further improves performance significantly.

Fig. 5 shows sample detections of the 3 methods on test images. The 4D-ISM detector is specifically designed to detect people in cluttered scenes with partial occlusions. Its drawback is that it tends to produce hypotheses even when little image evidence is available (image 3 and 4), which results in increased number of false positives. The HOG detector seems to have difficulties with the high variety in articulations and appearance present in out dataset. However, we should note that it is a multi-view detector designed to solve a more general problem than we consider here.

In addition to the high precision of our detector, we observe an improved scale estimation of the hypotheses as can be seen on the leftmost image of Fig. 5. We attribute this to the fact that the influence of imprecise scale-estimates of the local feature detector are reduced using local object parts.

3. Detection of Tracks in Image Sequences

The person detector just described provides hypotheses for position, scale, and body articulation in single frames based on the detection of individual body parts or limbs. To further improve the detection performance in image sequences several authors have proposed to incorporate temporal consistency among subsequent frames, typically using a simple motion model based on position and velocity of the person. In contrast, we propose to use a more expressive kinematic limb model thereby leveraging the articulated tracking literature (*e.g.* [5, 4, 21, 23, 19, 22]). Clearly, the expressiveness and the robustness of the kinematic limb model are crucial as it has to be powerful enough to reduce the number of false positives significantly, and at the same time robust enough to enable people detection in crowded scenes.

Given the image evidence $E = [E_1, E_2, \ldots, E_m]^T$ in a sequence of m subsequent frames, we would like to recover the positions $\mathbf{X}^{o*} = [\mathbf{x}_1^{o*}, \mathbf{x}_2^{o*}, \ldots, \mathbf{x}_m^{o*}]^T$ of the person as well as the configurations of the limbs in each frame $\mathbf{Y}^* = [\mathbf{y}_1^*, \mathbf{y}_2^*, \ldots, \mathbf{y}_m^*]^T$ with $\mathbf{y}_j^*$ denoting the recovered limb orientations in the j-th frame. Assuming independence of the detections in each frame, the posterior factorizes as:

$$\begin{aligned} p(\mathbf{Y}^*, \mathbf{X}^{o*}|E) &\propto p(\mathbf{Y}^*)p(\mathbf{X}^{o*})p(E|\mathbf{Y}^*, \mathbf{X}^{o*}) \quad (6) \\ &\propto p(\mathbf{Y}^*)p(\mathbf{X}^{o*}) \prod_{j=1}^{m} p(E_j|\mathbf{y}_j^*, \mathbf{x}_j^{o*}). \end{aligned}$$

$p(E_j|\mathbf{y}_j^*, \mathbf{x}_j^{o*})$ is the likelihood of the image evidence E_j,

[1] Available at www.mis.informatik.tu-darmstadt.de.

and is given by the detection model described in the previous section. $p(\mathbf{X}^{o*})$ corresponds to a prior of human body speed, which we model as a broad Gaussian. Probably the most interesting term is $p(\mathbf{Y}^*)$, which denotes the prior over the kinematic limb-motions, and in general is difficult to estimate reliably due to the high dimensionality of the pose space. Instead of modelling the pose dynamics directly in an high-dimensional space, several authors [23, 25, 22] have argued and shown that a low-dimensional representation is sufficient to approximate the dynamics. In the following we use a Gaussian process latent variable model (GPLVM) to obtain such a low-dimensional representation and discuss how it can be used to obtain reliable people detections in image sequences.

3.1. Gaussian process latent variable model

Let $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_m]^T$ be a sequence of D-dimensional observations (here describing the relative joint angles of body limbs). GPLVMs model the D-dimensional observation space as the output of D Gaussian processes with an input space of dimensionality q, where $q < D$. Each observation $\mathbf{y}_i$ is associated with a q-dimensional latent point $\mathbf{z}_i$. The likelihood of the observation sequence $\mathbf{Y}$ given the latent sequence $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_m]^T$ and model parameters θ is given by [13]:

$$p(\mathbf{Y}|\mathbf{Z}, \theta) = \prod_{i=1}^{D} \mathcal{N}(\mathbf{Y}_{:,i}|0, \mathbf{K}_\mathbf{z}), \tag{7}$$

where $\mathbf{Y}_{:,i}$ is the vector of values of feature i across all observations, and $\mathbf{K}_\mathbf{z}$ is the covariance matrix with elements given by a covariance function $k(\mathbf{z}_i, \mathbf{z}_j)$. In this paper we use a squared exponential covariance function augmented by Gaussian white noise. For a given $\mathbf{Y}$ we can find the positions of the latent points $\mathbf{Z}$ along with the model parameters θ by maximizing their likelihood from Eq. (7).

In addition to the low-dimensional latent representation of the data, GPLVMs provide a probabilistic mapping from the latent space to the observation space. One possibility to define a dynamic model in the latent space is to place a suitable prior on the elements of $\mathbf{Z}$. Such a prior can be given by a Gaussian process with time as input variable [14]. Given the sequence of points in time, $\mathbf{T} = [t_1, t_2, \ldots, t_m]^T$ at which the observations $\mathbf{Y}$ were made, the prior over $\mathbf{Z}$ is given by

$$p(\mathbf{Z}|\mathbf{T}) = \prod_{i=1}^{q} \mathcal{N}(\mathbf{Z}_{:,i}|0, \mathbf{K}_\mathbf{T}) \tag{8}$$

where $\mathbf{K}_\mathbf{T}$ is the covariance matrix of the time points. The covariance function in the time space can again be taken as squared exponential, which ensures smoothness of the trajectories.

We now combine this prior with the likelihood from Eq. (7), and maximize w.r.t. $\mathbf{Z}$ and θ. Fig. 6 shows the 2 dimensional latent space obtained by applying this model to 11 walking sequences of different subjects, each containing one complete walking cycle. Walking cycles in each sequence are manually aligned so that we can interpret the frame number in each sequence as phase of the walking cycle. This hierarchical approach to GPLVM dynamics has several advantages over the auto-regressive prior proposed in [25]. In particular, it allows us to evaluate the likelihood of a sequence of poses, even if the poses occurred at unequally spaced time intervals. This arises, *e.g.*, when the subject was occluded or not detected for several frames. Additionally, for a given pose the model allows us to hypothesize both successive and previous poses, which we use to produce good initial hypotheses for the whole image sequence from a few good detections.

3.2. Reconstruction of poses in short sequence

Given limb likelihoods and the hGPLVM prior, we can maximize Eq. (6) to find the best pose sequence. This is equivalent to jointly solving the inverse kinematics in each frame of the sequence under soft constraints given by limb likelihoods and similar to [9], except that in our case hints about limb positions are provided by detector instead of being manually given by the user. If we denote the training observations, their latent representation and model parameters by $\mathcal{M} = [\mathbf{Y}, \mathbf{T}, \mathbf{Z}, \boldsymbol{\theta}]$, the probability of the unknown pose sequence $\mathbf{Y}^*$, its latent representation $\mathbf{Z}^*$, and the person positions $\mathbf{X}^{o*}$ is given by

$$\begin{aligned} &p(\mathbf{Y}^*, \mathbf{X}^{o*}, \mathbf{Z}^*|\mathcal{M}, E, \mathbf{T}^*) \propto \\ &p(E|\mathbf{Y}^*, \mathbf{X}^{o*}) p(\mathbf{Y}^*|\mathbf{Z}^*, \mathcal{M}) p(\mathbf{Z}^*|\mathbf{T}^*, \mathcal{M}) p(\mathbf{X}^{o*}). \end{aligned} \tag{9}$$

The first term is the detection likelihood from single-frame detections (see Eq. (6)). The second term is given by

$$p(\mathbf{Y}^*|\mathbf{Z}^*, \mathcal{M}) = \prod_{i=1}^{D} p(\mathbf{Y}^*_{:,i}|\mathbf{Z}^*, \mathbf{Y}_{:,i}, \mathbf{Z}), \tag{10}$$

where $p(\mathbf{Y}^*_{:,i}|\mathbf{Z}^*, \mathbf{Y}_{:,i}, \mathbf{Z})$ is a Gaussian process prediction of the pose sequence given a sequence of latent positions. The third term is given by the dynamics prior on the latent space:

$$p(\mathbf{Z}^*|\mathbf{T}^*, \mathcal{M}) = \prod_{i=1}^{q} p(\mathbf{Z}^*_{:,i}|\mathbf{T}^*, \mathbf{Z}_{:,i}, \mathbf{T}). \tag{11}$$

In our formulation, detecting people in a series of m frames therefore corresponds to finding pose sequences $\mathbf{Y}^*$ and people positions $\mathbf{X}^{o*}$ that maximize Eq. (9). We use the following strategy to efficiently obtain such maxima: Each hypothesis obtained from the people detector from Sec. 2 contains an estimate of the person's position $\mathbf{x}^o$, limbs' positions $\mathbf{x}^i$ and articulation a. From these parameters we can

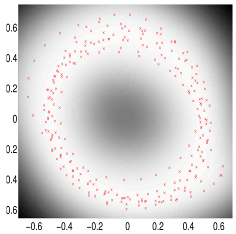


Figure 6. Representation of articulations in the latent space.

directly estimate both the limb-orientations $\mathbf{y}$, and position in the walking cycle t. Using those parameters and propagating them to neighboring frames using the kinematic limb model has proven to yield good initializations for optimization. In the experiments described below we use a sufficient but small number of detections in each frame to obtain initialization values for $\mathbf{Y}^*$, $\mathbf{T}^*$, and $\mathbf{X}^{o*}$, and then use a conjugate gradient method to find local maxima of Eq. (9). The gradients of the second, third, and fourth term in Eq. (9) can be computed analytically, while we use a finite difference approximation for gradient of $p(E|\mathbf{Y}^*, \mathbf{X}^{o*})$. As the experiments show, this enables us to efficiently obtain people detections in image sequences.

Quantitatively, Fig. 7(b) shows how the extracted tracklets lead to increased precision and recall compared to the detector alone (note that the recall does not reach 1 in either case due to the use of non-maxima suppression.)

3.3. Optimal track selection based on overlapping tracklets

The optimization procedure just described is suited to reliably detect people in short frame sequences. We found that in order to reconstruct tracks of people over longer periods of time, it is more reliable to merge hypotheses from different short tracklets rather than increasing the length of the tracklets itself. First, we compute overlapping fixed-length tracklets ($m = 6$) starting at every frame of the sequence. As tracklet optimization relies on good initialization, the use of overlapping sequences ensures that each strong detection is used multiple times for initialization.

We then exploit that every frame is overlapped by several different tracklets (including ones with different starting frames), which provide competing hypotheses that explain the same image evidence. In principle it would be possible to choose the best sequence of hypotheses using their joint posterior (*i.e.*, an extension of Eq. (9)). This is, however, computationally prohibitive since the large state-space of all possible combinations of hypotheses cannot be searched efficiently without making simplifying assumptions. Instead, we select hypotheses using pairwise relations between them by introducing a first-order Markov assumption on the hypothesis sequence.

Let the length of the complete image sequence be equal to M. We denote the set of all hypotheses obtained from individual tracklets in frame j by $\mathbf{h}^j = [h_1^j, \ldots, h_{n_j}^j]$. We will call the track given by a set of hypotheses $\mathcal{H} = [h_{j_1}^1, \ldots, h_{j_M}^M]$ optimal if it maximizes the joint sequence probability according to the hidden Markov model:

$$p(\mathcal{H}) = p_{img}(h_{j_1}^1) \prod_{k=2}^{M} p_{img}(h_{j_k}^k) p_{trans}(h_{j_k}^k, h_{j_{k-1}}^{k-1}). \quad (12)$$

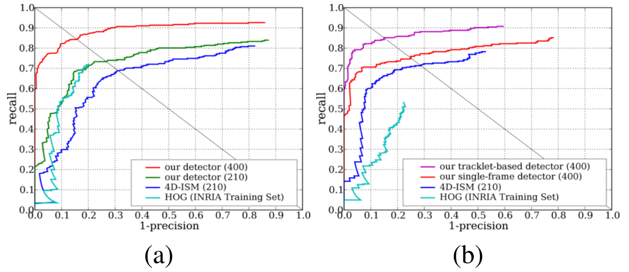
In this expression $p_{img}(h_{j_k}^k)$ is computed using the people detection model from Sec. 2. The transition probability is p_{trans} is given by

$$\begin{aligned} p_{trans}(h_{j_k}^k, h_{j_{k-1}}^{k-1}) &= p_{dynamic}(h_{j_k}^k, h_{j_{k-1}}^{k-1}) \cdot \\ & \quad p_{app}(h_{j_k}^k, h_{j_{k-1}}^{k-1}), \end{aligned} \quad (13)$$

where $p_{dynamic}(\cdot, \cdot)$ is our dynamical model consisting of Gaussian position dynamics and the GPLVM articulation dynamics, and $p_{app}(\cdot, \cdot)$ is computed using an appearance model of the hypothesis, which can be based on color histograms of person parts, oriented edge features, or any other appearance description. We use color histograms extracted from the detected parts, and use the Bhattacharyya distance to model the appearance compatibility.

The optimal sequence can be efficiently found by maximizing Eq. (12) using the Viterbi algorithm. If a given image sequence contains only one person that is never occluded, the proposed approach is able to reconstruct its track from the individual tracklets.

For the case of more complex image sequences we have adopted the following strategy that has proven to be quite effective (see Sec. 4 for results): Let i be the number of the current frame in the sequence (in the beginning $i = 1$). We proceed by iteratively computing the optimal track starting with i. At the nth iteration of the Viterbi algorithm, we compute the transition probabilities between the hypotheses of the optimal track at frame $i + n$ and each of the hypotheses in the frame $i + n + 1$. If the person either walks out of the image or becomes heavily occluded, all of the transition probabilities will be low, which means that we can end the track. In that case all its hypotheses are removed from the sets of hypotheses $\mathbf{h}^j$, $j = i, \ldots, i + n$. We then repeat the procedure again starting from frame i until $\mathbf{h}^i$ becomes empty. In this case we set $i = i + 1$ and repeat the process. As a result of this iterative computation we obtain a set of tracks with hypotheses that are consistent in both motion and articulation. To connect such tracks across long-term occlusions we again use the appearance of the person as well as a coarse motion model to decide if two tracks correspond to the same person. The appearance model is the same as used for modeling the appearance of individual hypotheses. For the motion model we only require tracks to have consistent movement direction (i.e. left or right). In practice, we found that even such a simplistic method is



(a) (b)

Figure 7. Comparison of our pedestrian detector with 4D-ISM [20] and HOG [3] on (a) the “TUD-Pedestrians” and (b) “TUD-Campus” datasets. Numbers in parenthesis indicate number of training images.

Dataset	HOG		4D-ISM		single-frame		tracklets	
TUD-Ped	0.53	-	0.28	0.68	0.81	0.84	-	-
TUD-Camp	0.22	-	0.6	0.71	0.7	0.75	0.82	0.85

Table 1. Recall of 4D-ISM, HOG, and our detectors at precision equal to 0.9 and at equal error rate on “TUD-Pedestrians” and “TUD-Campus” datasets.

sufficient since only few possible tracks are available after the initial detection stage.

4. Experiments

We evaluate our approach both quantitatively and qualitatively. In the first experiment, we compare the detection performance of the single-frame person detector proposed in Sec. 2 with the tracklet-based detector. Tracklet-based detections are obtained by first detecting tracklets in the image sequence as described in Sec. 3.2, grouping together hypotheses from all tracklets corresponding to a particular image of the sequence, and performing non-maxima suppression on this set of hypotheses. In this process the score of a hypothesis is set to score of the corresponding tracklet, which is given by Eq. 9.

The comparison of both detectors is done on a sequence with multiple full and partial occlusions. Fig. 9 shows several example images from the sequence. Note that in such cluttered sequences ground truth annotation is difficult as it is unclear how to decide when a partially or fully occluded person should be included. In order to have a fair evaluation for the single-frame person detector we decided to annotate people when they are at least 50% visible. The quantitative comparison of the single-frame detector and the tracklet-based detector is given in Fig. 7. The tracklet-based detector improves the precision considerably. Fig. 8 shows sample detections. As can be seen, the single-frame detector obtains false positives on the background (images 1, 2, and 3), whereas the tracklet-based detector can successfully filter those false-positives. At the same time the tracklet-based detector is capable of detecting several partially occluded people (*e.g.* in image 2 and 4) that cannot be detected in a single frame alone.

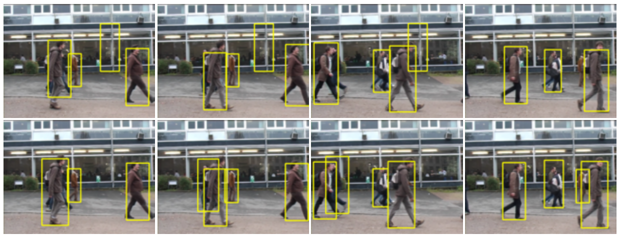


Figure 8. Examples of detector hypotheses (top row) and tracklet hypotheses (bottom row) at equal error rate on the “TUD-Campus” dataset.

In the second experiment we evaluate the tracks produced by our system on the sequence from the first experiment and an additional sequence with significantly larger number of people. Example snapshots from the resulting tracks are shown in figures 9 and 10 respectively. Clearly, in any single frame a significant number of people is detected and their limb-configuration is correctly inferred. Interestingly, we obtain tracks for nearly all people in these sequences, and in particular in sequence 1 we obtain tracks for all people even though some of them become fully occluded over significant time intervals. Quite importantly, on this sequence we can also differentiate between hypotheses of two people walking side by side, with one person occluding the other most of the time. The complete videos of both sequences can be found in the supplementary material.

5. Conclusion

This paper proposed a novel method capable of detecting and tracking people in cluttered real-world scenes with many people and changing backgrounds. For this the paper extended a state-of-the-art pedestrian detector to an articulation and limb-based detection approach, which outperforms the state-of-the-art on single frame person detection. A dynamic limb-model based on a hierarchical Gaussian process latent variable model is used to further improve people-detection by people-tracklet detection in image sequences. Those tracklets are then used to enable people-tracking in complex scenes with many people and long-term occlusions.

In the future we will extend the proposed approach using a 3D limb model to allow people-detection from arbitrary viewpoints and across multiple cameras.

Acknowledgements: The authors are very thankful to Neil Lawrence for making his GPLVM toolbox available online and to Mario Fritz for helpful suggestions concerning the color model. Mykhaylo Andriluka gratefully acknowledges a scholarship from DFG GRK “Cooperative, Adaptive and Responsive Monitoring in Mixed Mode Environments”.

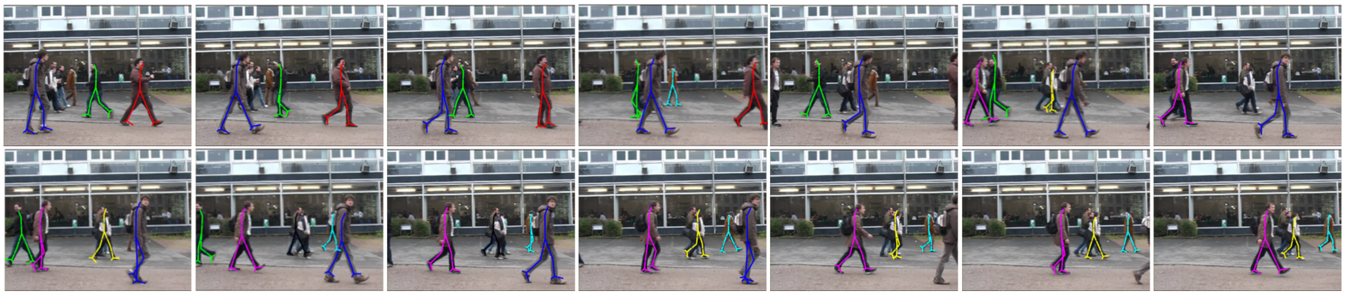


Figure 9. Detection and tracking on “TUD-Campus” dataset (see supplementary material).

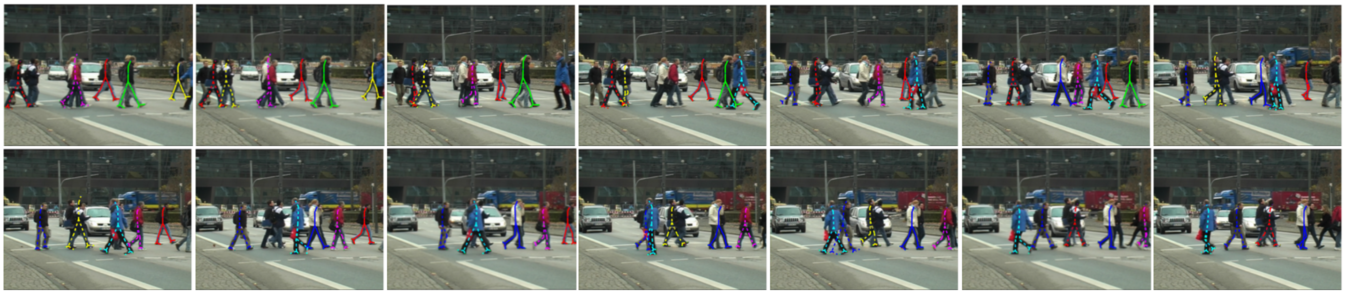


Figure 10. Detection and tracking on “TUD-Crossing” dataset (see supplementary material).

References

[1] S. Avidan. Ensemble tracking. *PAMI*, 29:261–271, 2007.

[2] S. Belongie, J. Malik, and J. Puzicha. Shape context: A new descriptor for shape matching and object recognition. *NIPS*2000*.

[3] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. *CVPR 2005*.

[4] D. Demirdjian, L. Taycher, G. Shakhnarovich, K. Grauman, and T. Darrell. Avoiding the ”streetlight effect”: Tracking by exploring likelihood modes. *ICCV 2005*.

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

[6] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *IJCV*, 61:55–79, 2007.

[7] A. Fossati, M. Dimitrijevic, V. Lepetit, and P. Fua. Bridging the gap between detection and tracking for 3D monocular video-based motion capture. *CVPR 2007*.

[8] H. Grabner and H. Bischof. On-line boosting and vision. *CVPR 2006*.

[9] K. Grochow, S. L. Martin, A. Hertzmann, and Z. Popovic. Style-based inverse kinematics. *ACM SIGGRAPH*, 2004.

[10] S. Ioffe and D. Forsyth. Human tracking with mixtures of trees. *ICCV 2001*.

[11] M. Isard and J. MacCormick. BraMBLe: A Bayesian multiple-blob tracker. *ICCV 2001*.

[12] X. Lan and D. P. Huttenlocher. Beyond trees: Common-factor models for 2d human pose recovery. *ICCV 2005*.

[13] N. D. Lawrence. Probabilistic non-linear principal component analysis with Gaussian process latent variable models. *J. Mach. Learn. Res.*, 6:1783–1816, 2005.

[14] N. D. Lawrence and A. J. Moore. Hierarchical Gaussian process latent variable models. *ICML 2007*.

[15] B. Leibe, K. Schindler, and L. Van Gool. Coupled detection and trajectory estimation for multi-object tracking. *ICCV 2007*.

[16] B. Leibe, E. Seemann, and B. Schiele. Pedestrian detection in crowded scenes. *CVPR 2005*.

[17] K. Mikolajczyk and C. Schmid. Scale and affine invariant interest point detectors. *IJCV*, 60:63–86, 2004.

[18] K. Okuma, A. Taleghani, N. De Freitas, J. J. Little, and D. G. Lowe. A boosted particle filter: Multitarget detection and tracking. *ECCV 2004*.

[19] D. Ramanan, D. A. Forsyth, and A. Zisserman. Tracking people by learning their appearance. *PAMI*, 29:65–81, 2007.

[20] E. Seemann and B. Schiele. Cross-articulation learning for robust detection of pedestrians. *DAGM*, 2006.

[21] L. Sigal and M. J. Black. Measure locally, reason globally: Occlusion-sensitive articulated pose estimation. *CVPR 2006*.

[22] C. Sminchisescu, A. Kanaujia, and D. N. Metaxas. BM^{3}E: Discriminative density propagation for visual tracking. *PAMI*, 29:2030–2044, 2007.

[23] R. Urtasun, D. J. Fleet, and P. Fua. 3D people tracking with Gaussian process dynamical models. *CVPR 2006*.

[24] P. Viola and M. Jones. Robust real-time face detection. *IJCV*, 57:137–164, 2004.

[25] J. M. Wang, D. J. Fleet, and A. Hertzmann. Gaussian process dynamical models. *NIPS*2005*.

[26] C. K. I. Williams and M. Allan. On a connection between object localization with a generative template of features and pose-space prediction methods. Technical Report EDI-INF-RR-0719, University of Edinburgh, 2006.

[27] 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, 2007.