Pictorial Structures Revisited: People Detection and Articulated Pose Estimation

Mykhaylo Andriluka, Stefan Roth, and Bernt Schiele
Department of Computer Science, TU Darmstadt

Abstract

Non-rigid object detection and articulated pose estimation are two related and challenging problems in computer vision. Numerous models have been proposed over the years and often address different special cases, such as pedestrian detection or upper body pose estimation in TV footage. This paper shows that such specialization may not be necessary, and proposes a generic approach based on the pictorial structures framework. We show that the right selection of components for both appearance and spatial modeling is crucial for general applicability and overall performance of the model. The appearance of body parts is modeled using densely sampled shape context descriptors and discriminatively trained AdaBoost classifiers. Furthermore, we interpret the normalized margin of each classifier as likelihood in a generative model. Non-Gaussian relationships between parts are represented as Gaussians in the coordinate system of the joint between parts. The marginal posterior of each part is inferred using belief propagation. We demonstrate that such a model is equally suitable for both detection and pose estimation tasks, outperforming the state of the art on three recently proposed datasets.

1. Introduction and Related Work

Both people detection and human pose estimation have a large variety of applications such as automotive safety, surveillance, and video indexing. The goal of this paper is to develop a generic model for human detection and pose estimation that allows to detect upright people (*i.e.*, pedestrians [12]), as well as highly articulated people (*e.g.*, in sports scenes [15]), and to estimate their poses. Our model should also enable upper body detection and pose estimation [6], *e.g.*, for movie indexing. The top performing methods for these three scenarios do currently not share the same architecture, nor are components necessarily similar either. Here, we present a generic approach that allows for both human detection and pose estimation thereby addressing the above mentioned scenarios in a single framework. Due to its careful design the proposed approach outperforms recent work on three challenging datasets (see Fig. 1 for examples).

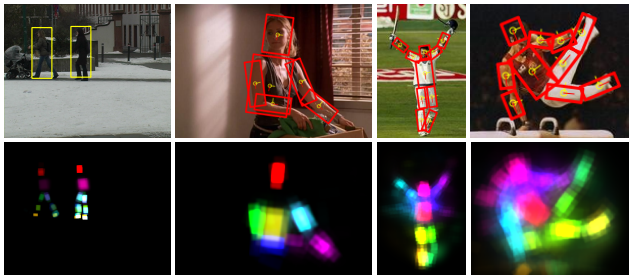


Figure 1. **Example results** (from left to right): Pedestrian detection, upper-body pose estimation, and full body pose estimation (3rd and 4th column) using our method. *Bottom:* Part posteriors.

Our work builds upon the *pictorial structures* model [4, 6, 15], which is a powerful and general, yet simple generative body model that allows for exact and efficient inference of the part constellations. We also build upon *strong part detectors* [1, 13, 24], which have shown to enable object and people detection in challenging scenes, but have not yet proven to enable state-of-the-art articulated pose estimation. While previous work has either focused on strong part detectors or on powerful body models, our work combines the strengths of both.

The original pictorial structures approach of Felzenszwalb and Huttenlocher [4] is based on a simple appearance model requiring background subtraction, which renders it inappropriate for the scenes considered here. In [18] the approach has been demonstrated to work without background substraction by relying on a discriminative appearance model, but while still using rather simple image features (Gaussian derivatives). More powerful part templates are extracted in [15] using an iterative parsing approach. This was later extended by [6], which furthermore integrates features from an automatic foreground segmentation step to improve performance, which we do not require here. Both approaches iteratively build more powerful detectors to reduce the search space of valid articulations, but use relatively weak edge cues at the initial detection stage.

Our approach, on the other hand, uses strong generic part detectors that do not require iterative parsing or other ways of reducing the search space, other than of course an articulated body model. In particular, we compute dense appearance representations based on shape context descriptors

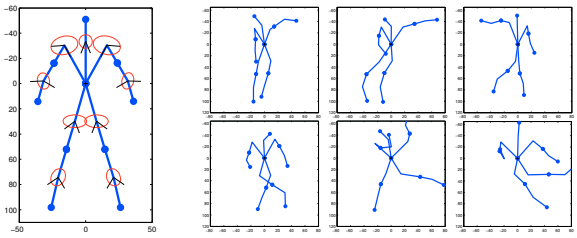


Figure 2. *(left)* **Kinematic prior** learned on the multi-view and multi-articulation dataset from [15]. The mean part position is shown using blue dots; the covariance of the part relations in the transformed space is shown using red ellipses. *(right)* **Several independent samples** from the learned prior (for ease of visualization given fixed torso position and orientation).

[14], and use AdaBoost [7] to train discriminative part classifiers. Our detectors are evaluated densely and are bootstrapped to improve performance. Strong detectors of that type have been commonplace in the pedestrian detection literature [1, 12, 13, 24]. In these cases, however, the employed body models are often simplistic. A simple star model for representing part articulations is, for example, used in [1], whereas [12] does not use an explicit part representation at all. This precludes the applicability to strongly articulated people and consequently these approaches have been applied to upright people detection only.

We combine this discriminative appearance model with a generative pictorial structures approach by interpreting the normalized classifier margin as the image evidence that is being generated. As a result, we obtain a generic model for people detection and pose estimation, which not only outperforms recent work in both areas by a large margin, but is also surprisingly simple and allows for exact and efficient inference.

More related work: Besides the already mentioned related work there is an extensive literature on both people (and pedestrian) detection, as well as on articulated pose estimation. A large amount of work has been advocating strong body models, and another substantial set of related work relies on powerful appearance models.

Strong body models have appeared in various forms. A certain focus has been the development of non-tree models. [17] imposes constraints not only between limbs on the same extremity, but also between extremities, and relies on integer programming for inference. Another approach incorporate self-occlusion in a non-tree model [8]. Either approach relies on matching simple line features, and only appears to work on relatively clean backgrounds. In contrast, our method also works well on complex, cluttered backgrounds. [20] also uses non-tree models to improve occlusion handling, but still relies on simple features, such as color. A fully connected graphical model for representing articulations is proposed in [2], which also uses discriminative part detectors. However, the method has several restrictions, such as relying on absolute part orientations, which makes it applicable to people in upright poses only. Moreover, the fully connected graph complicates inference. Other work has focused on discriminative tree models [16, 18], but due to the use of simple features, these methods fall short in terms of performance. [25] proposes a complex hierarchical model for pruning the space of valid articulations, but also relies on relatively simple features. In [5] discriminative training is combined with strong appearance representation based on HOG features, however the model is applied to detection only.

Discriminative part models have also been used in conjunction with generative body models, as we do here. [11, 21], for example, use them as proposal distributions ("shouters") for MCMC or nonparametric belief propagation. Our paper, however, directly integrates the part detectors and uses them as the appearance model.

2. Generic Model for People Detection and Pose Estimation

To facilitate reliable detection of people across a wide variety of poses, we follow [4] and assume that the body model is decomposed into a set of parts. Their configuration is denoted as $L = \{\mathbf{l}_0, \mathbf{l}_1, \ldots, \mathbf{l}_N\}$, where the state of part i is given by $\mathbf{l}_i = (x_i, y_i, \theta_i, s_i)$. x_i and y_i is the position of the part center in image coordinates, θ_i is the absolute part orientation, and s_i is the part scale, which we assume to be relative to the size of the part in the training set.

Depending on the task, the number of object parts may vary (see Figs. 2 and 3). For upper body detection (or pose estimation), we rely on 6 different parts: head, torso, as well as left and right lower and upper arms. In case of full body detection, we additionally consider 4 lower body parts: left and right upper and lower legs, resulting in a 10 part model. For pedestrian detection we do not use arms, but add feet, leading to an 8 part model.

Given the image evidence D, the posterior of the part configuration L is modeled as $p(L|D) \propto p(D|L)p(L)$, where $p(D|L)$ is the likelihood of the image evidence given a particular body part configuration. In the pictorial structures approach $p(L)$ corresponds to a kinematic tree prior. Here, both these terms are learned from training data, either from generic data or trained more specifically for the application at hand. To make such a seemingly generic and simple approach work well, and to compete with more specialized models on a variety of tasks, it is necessary to carefully pick the appropriate prior $p(L)$ and an appropriate image likelihood $p(D|L)$. In Sec. 2.1, we will first introduce our generative kinematic model $p(L)$, which closely follows the pictorial structures approach [4]. In Sec. 2.2, we will then introduce our discriminatively trained appearance model $p(D|L)$.

Given such a model, we estimate articulated poses by finding the most probable location for each part given the image evidence through maximizing the marginal posterior $p(\mathbf{l}_i|D)$. In case of multiple people this directly generalizes to finding the modes of the posterior density.

To address the problem of people detection, we also rely on an articulated body model (like *e.g.*, [1, 15]) to be able to cope with the large variety of possible body poses. To that end we first compute the marginal distribution of the position of the torso, and then use its modes to deterministically predict the positions of the detection bounding boxes.

2.1. Kinematic tree prior

The first important component in our pictorial structures approach is the prior $p(L)$, which encodes probabilistic constraints on part configurations. A common source of such constraints are kinematic dependencies between parts. Mapping the kinematic structure on a directed acyclic graph (DAG), the distribution over configurations can be factorized as

$$p(L) = p(\mathbf{l}_0) \prod_{(i,j)\in E} p(\mathbf{l}_i|\mathbf{l}_j), \qquad (1)$$

where we let E denote the set of all directed edges in the kinematic tree and assign $\mathbf{l}_0$ to be the root node (torso).

It is certainly possible to incorporate action specific constraints into the prior, and to combine them with articulation dynamics to enable tracking, as is done for example in [10, 23]. However, we omit such extensions, since they would restrict the applicability of the model to rather specific scenarios.

Part relations. To complete the specification of the prior, we have to specify the various components of Eq. (1). The prior for the root part configuration $p(\mathbf{l}_0)$ is simply assumed to be uniform, to allow for a wide range of possible configurations. The part relations are modeled using Gaussian distributions (*c.f.* [4, 16]), which allow for efficient inference (see below). This may seem like a significant limitation, for example as the distribution of the forearm position given the upper arm position intuitively has a semi-circular rather than a Gaussian shape. It was pointed out in [4] that while such a distribution is not Gaussian in the image coordinates, it is possible to transform it to a different space, in which the spatial distribution between parts is again well captured by a Gaussian distribution. More specifically, to model $p(\mathbf{l}_i|\mathbf{l}_j)$, we transform the part configuration $\mathbf{l}_i = (x_i, y_i, \theta_i, s_i)$ into the coordinate system of the joint between the two parts using the transformation:

$$T_{ji}(\mathbf{l}_i) = \begin{pmatrix} x_i + s_i d_x^{ji} \cos\theta_i - s_i d_y^{ji} \sin\theta_i \\ y_i + s_i d_x^{ji} \sin\theta_i + s_i d_y^{ji} \cos\theta_i \\ \theta_i + \tilde{\theta}_{ji} \\ s_i \end{pmatrix}. \qquad (2)$$

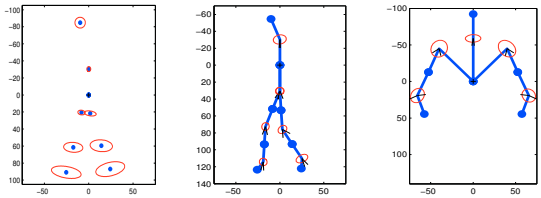


Figure 3. **Priors on the part configurations** (left to right): Pedestrian detection (star *vs*. tree model) and upper body detection.

Here, $d^{ji} = (d_x^{ji}, d_y^{ji})^T$ is the mean relative position of the joint between parts i and j in the coordinate system of part i and $\tilde{\theta}_{ji}$ is the relative angle between parts. Then the part relation is modeled as a Gaussian in the transformed space:

$$p(\mathbf{l}_i|\mathbf{l}_j) = \mathcal{N}(T_{ji}(\mathbf{l}_i)|T_{ij}(\mathbf{l}_j), \Sigma^{ji}), \qquad (3)$$

where T_{ij} is the transformation that maps position of parent part $\mathbf{l}_j$ to the position of the joint between parts i and j, and Σ^{ji} is the covariance between the parts that we learn from data, which determines the stiffness of the joint. Moreover, we need to learn the mean relative joint position d^{ji}. Both d^{ji} and Σ^{ji} can be learned in a quite straightforward way using maximum likelihood estimation. One important thing to note is that this corresponds to a so-called "loose limbed" model (*c.f.* [21]), because the limbs do not rigidly rotate around the joints. Instead, the parts are only loosely attached to the joint by means of the Gaussian distribution from Eq. (3), which helps reduce brittle behavior. In our experiments, we found this simple and efficient procedure to work much better than the non-parametric part relation model used in [15].

Learned prior. Figure 2 shows the prior learned from the multi-view and multi-articulation people dataset from [15], which includes people involved in large variety of activities ranging from simple walking to performing acrobatic exercises. Samples from this model (see Fig. 2) exhibit a large variety of poses. Fig. 3 also shows priors learned on the TUD-Pedestrians dataset [1], which contains upright pedestrians in street scenes, and from the "Buffy" dataset [6], which contains upper body configurations in TV footage.

2.2. Discriminatively trained part models

The other important component in our formulation is the likelihood $p(D|L)$ of the image evidence D given the part configuration L. We rely on part specific appearance models, each of which will result in a part evidence map $\mathbf{d}_i$ that reports the evidence for part i for each possible position, scale, and rotation. To unify this discriminative appearance model with the generative body model, we assume that the part evidence maps are generated based on knowing the true body configuration L. Assuming that the different part evidence maps are conditionally independent given the configuration L, and that the part map $\mathbf{d}_i$ for part i only depends

on its own configuration $\mathbf{l}_i$, the likelihood simplifies as:

$$p(D|L) = \prod_{i=0}^{N} p(\mathbf{d}_i|L) = \prod_{i=0}^{N} p(\mathbf{d}_i|\mathbf{l}_i). \quad (4)$$

While this is clearly a simplifying assumption, it is justifiable as long as parts do not occlude each other significantly (*c.f.* [4]). Moreover, this enables efficient and exact inference and leads to very competitive experimental results. As a consequence of Eq. (4) the posterior over the configuration of parts factorizes as:

$$p(L|D) \propto p(\mathbf{l}_0) \cdot \prod_{i=0}^{N} p(\mathbf{d}_i|\mathbf{l}_i) \cdot \prod_{(i,j)\in E} p(\mathbf{l}_i|\mathbf{l}_j), \quad (5)$$

Challenges. There are a number of considerations that drive the choices behind our discriminative appearance model. Our aim is to detect people in unconstrained environments and arbitrary poses. As the search space for all possible poses is often very large, search space reduction can be an important component of a successful approach [6]. As has been argued before [22], using discriminatively learned detectors allows to reduce the search space for the generative model significantly thereby enabling not only efficient but also effective inference in challenging real world scenes. Following this avenue, we rely on a part-specific discriminative appearance model to effectively reduce the search space as much as possible. In a similar vein, it is important to avoid prefiltering the possible part locations at the part detection stage, and to postpone the final decision until evidence from all body parts is available. Therefore we densely evaluate the search space and consider all possible part positions, orientations, and scales, which is in contrast to bottom-up appearance models (*e.g.* [1, 13]) based on a sparse set of local features. We believe that dense sampling is better suited for detecting body parts, especially in cases of low contrast and partial occlusion.

Boosted part detectors. Our discriminative part detectors densely sample a variant of the shape context descriptor initially proposed in [14] and previously used for pedestrian detection [19]. In this descriptor the distribution of locally normalized gradient orientations is captured in a log-polar histogram. The log-polar binning is especially suited for our task, since it is tolerant to small changes in the rotation of the body parts. In our experiments we use 12 bins for the location and 8 bins for the gradient orientation, which results in a 96 dimensional descriptor. We ignore the sign of the gradient as we found this to improve generalization.

The feature vector used for classification is obtained by concatenating all shape context descriptors whose centers fall inside of the part bounding box, so that some of the feature vector dimensions also capture the surrounding context. During detection all possible positions, scales, and orientations are scanned in a sliding window fashion. To predict the presence of a part, we train an AdaBoost classifier [7] with simple decision stumps that consider whether one of the log-polar histogram bins of the feature vector is above a threshold. Denoting the feature vector by $\mathbf{x}$, the stump with index t is given by $h_t(\mathbf{x}) = \text{sign}(\xi_t(x_{n(t)} - \varphi_t))$, where φ_t is a threshold, $\xi_t \in \{-1, +1\}$, and $n(t)$ is index of the histogram bin chosen by the stump. Training of the AdaBoost classifier proceeds as usual yielding a strong classifier $H_i(\mathbf{x}) = \text{sign}\left(\sum_t \alpha_{i,t} h_t(\mathbf{x})\right)$ for each part i. Here, $\alpha_{i,t}$ are the learned weights of the weak classifiers.

To integrate the discriminative classifiers into the generative probabilistic framework described above, it is necessary to give a probabilistic meaning to the classifier outputs. For that we interpret the normalized classifier margin as the likelihood:

$$\tilde{p}(\mathbf{d}_i|\mathbf{l}_i) = \max\left(\frac{\sum_t \alpha_{i,t} h_t(\mathbf{x}(\mathbf{l}_i))}{\sum_t \alpha_{i,t}}, \varepsilon_0\right) \quad (6)$$

where $\mathbf{x}(\mathbf{l}_i)$ denotes the feature for the part configuration $\mathbf{l}_i$, and ε_0 is a small positive constant, which makes the model more robust to part occlusions and missing parts. In our experiments we set $\varepsilon_0 = 10^{-4}$. As we show in Sec. 3 this simple pseudo-probability works quite well in practice.

Training. At the training stage, each annotated part is scaled and rotated to a canonical pose prior to learning. Note that this in-plane rotation normalization significantly simplifies the classification task. Additionally, we extend the training set by adding small scale, rotation, and offset transformations to the original images. The negative feature vectors are obtained by uniformly sampling them from the image regions outside of the object bounding box. After the initial training, the classifiers are re-trained with a new negative training set that has been augmented with false positives produced by the initial classifier. This is commonly referred to as bootstrapping. We have found that bootstrapping is essential for obtaining good performance with our discriminative part detectors (*c.f.* Fig. 4(c)).

2.3. Exact model inference

An important property of such a tree-based model is that optimal inference is tractable. Specifically, we could compute the globally optimal body configuration by doing MAP inference using the max-product algorithm [4]. Moreover, we can compute exact marginal distributions using the sum-product algorithm [9], which we employ here, because we require marginals for pedestrian detection.

To that end, we interpret the underlying directed graphical model as a factor graph, and apply standard factor graph belief propagation (sum-product).

An important observation is that if the part dependencies are modeled using Gaussian distributions, then expensive summations necessary in the sum-product algorithm can be

efficiently computed using Gaussian convolutions. However, care has to be taken when doing so, as the part relations in our model are Gaussian not in the image space, but rather in the transformed space of the joint. To apply the efficient algorithm nonetheless, we rely on the approach suggested in [4] and transform the messages into the coordinate system of the joint using Eq. (2), then apply Gaussian convolutions there, and finally transform the result into the coordinate system of the target part, which is possible since the transformation from the part position to the position of the joint is invertible. These computations are especially efficient if the Gaussian distribution in the transformed space is separable.

3. Experiments

We evaluate our model on three related tasks of increasing complexity: Pedestrian detection, upper body pose estimation, and multi-view full body pose estimation. For each task we use publicly available datasets and directly compare to the respective methods designed to work specifically on each of the tasks. In all experiments we discretize the part configurations in the same way. The discretization step is 1 pixel for the position in the image, 15 degrees for the part orientation, and 0.1 for the object scale.

3.1. Pedestrian detection

To detect pedestrians, we compute the marginal distribution of the torso location and use that to predict the pedestrian's bounding box. To deal with multiple peaks in the posterior corresponding to the same detection hypothesis, we perform a non-maximum suppression step. For each detection we remove all detections with smaller probability and more than 50% cover and overlap.

We use two datasets to evaluate different aspects of our model. The **TUD-Pedestrians** dataset [1] contains 250 images with 311 pedestrians (mostly side-views) with large variability in clothing and articulation. The training set contains 400 images. We created a new dataset called **TUD-UprightPeople**, which contains images from TUD-Pedestrians and additional images taken under various illumination conditions. This new dataset is used to evaluate different aspects of our model. The dataset contains 435 images with one person per image. We compare our approach to previous work on both datasets.

In all pedestrian detection experiments we use the 400 training images provided with TUD-Pedestrians to train the part detectors. We used two different priors on the part configuration: (1) A star prior in which all parts are connected directly to the center part; and (2) a kinematic tree prior (both are shown in Fig. 3).

To start our discussion consider Fig. 4(d), where we compare our approach (using 8 body parts) to results from the literature on the TUD-Pedestrians dataset. We use 8 parts either with the star prior or with the kinematic tree prior, both estimated on the training images of TUD-Pedestrians based on part annotations.

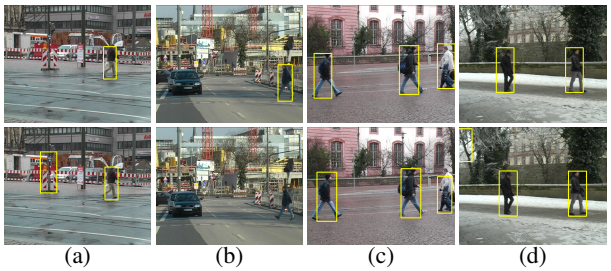
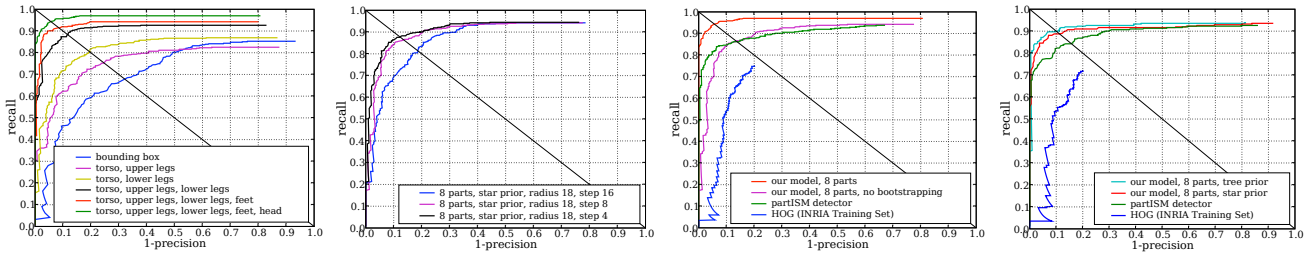


Figure 5. **Several examples of detections** at equal error rate obtained with our model (8 parts and tree prior, top) and partISM (bottom) on the "TUD-Pedestrians" dataset.

While the tree-based prior slightly outperforms the star prior, both outperform the partISM-model [1] by 4 and 5 % equal error rate (EER) respectively. They are also significantly better than the publicly available HOG binary [3], which however needs less supervision since it does not require part annotations during training. Similarly, Fig. 4(c) shows the same relative performance on the TUD-UprightPeople dataset. We attribute the improved performance over partISM to our dense part representation and to the discriminatively learned appearance model. PartISM, in contrast, uses a generative appearance model on sparse interest points. Fig. 5 shows sample detections of our model and partISM. Our model is flexible enough to capture diverse articulations of people as for example in Fig. 5(b), (c) and (d), which typically are problematic for monolithic detectors. However since our model is build on top of discriminative classifiers, it can avoid false positives in the background, which plague the generative partISM model (*e.g.*, Fig. 5(a) and (d)).

To gain more insight into the role of the different components, we conducted a series of experiments on the TUD-UprightPeople dataset. Here, we report results for the star prior only, as it allows to arbitrarily add and remove body parts from the model. Fig. 4(a) shows the influence of different numbers of parts on the detection performance. As expected, using more body parts generally improves the detection performance. We also compare to a monolithic detector that consists of single part defined by the person's bounding box (as is typical, *e.g.*, for the HOG detector [3]). This monolithic detector did not perform well even compared to detectors with as few as 3 parts.

In Fig. 4(b) we evaluate how the density of sampling the local features affects performance. In both cases the distance between evaluated part positions is kept the same, but the distance between features included in the feature vector presented to the AdaBoost classifier is set to 4, 8 and 16 pixels. The denser version produces consistently better results



(a) (b) (c) (d)

Figure 4. **Pedestrian detection results:** Performance of our model with different number of parts (a) and different step sizes between local features (b) on the "TUD-UprightPeople" dataset. Comparison with previously proposed approaches partISM [1] and HOG [3] on the "TUD-UprightPeople" dataset (c) and "TUD-Pedestrians" dataset (d). Note that denser sampling of local features (b) and bootstrapping (c) result in improvements of 6% and 8% EER respectively.

and outperforms the sparser one by approximately 6% EER. This confirms our intuition that an over-complete encoding of information at different spatial resolutions is important for classification. Note that the overall performance difference between Fig. 4(a) and (b) is due to the fact that we did not perform bootstrapping on the part detectors in 4(b).

3.2. Upper body pose estimation

To evaluate our method on the task of upper-body pose estimation, we use the recently proposed **Buffy** dataset [6], where the task is to estimate positions of torso, head, left and right forearm, and left and right upper arm in still frames extracted from 3 episodes of the popular TV show "Buffy the Vampire Slayer". This is very challenging due to large variability of poses, varying and loose fitting clothing, as well as strongly varying illumination. Due to the complexity of the task, the previously proposed approach [6] used multiple stages to reduce the search space of admissible poses. In particular, they perform an additional automatic foreground/background separation step based on 'GrabCut', and use spatio-temporal constraints. In our approach we directly estimate the pose from images without any search space pruning. Nonetheless, one can think of the discriminative part models we use as a form of pruning.

Ferrari *et al*. [6] report quantitative pose estimation results only for the single-frame detector (no temporal coherency) on a subset of people that have been correctly localized with a weak object detector, which is used for prefiltering. To facilitate comparison, we used the same weak detector (using the implementation made available by the authors of [6]), and only estimated the upper body pose in the image regions around correctly localized persons. Table 1 gives the detection results for each of the 6 upper body parts, along with the overall detection performance, which are measured using the same criterion as in [6]. A body part is considered correctly localized if the endpoints of its segment lie within 50% of the ground-truth segment length from their true positions.

We consider 3 different cases and report the localization performance (1) of the boosted part detectors alone; (2) of our full model with a generic upper body prior trained on the data used in Sec. 3.3 (*c.f*. Fig. 2); and (3) of our full model with a specialized front/back-view prior. Latter has been estimated on episode 4 of the "Buffy" dataset, which was not used for evaluation. In either case we used generic part detectors trained on the 100 training images of the "Iterative Image Parsing" dataset from [15].

As the results in Tab. 1 show, our method significantly outperforms the approach of [6] (71.3% *vs*. 57.9% average localization performance), even when using a generic body model. Yet in contrast to [6], we do not require separate foreground/background segmentation or use color features. The application specific front/back-view prior improves performance even further, albeit only slightly. It is also noteworthy that the part detectors alone, while powerful, do not perform well on this dataset, especially those of the arms. On one hand this illustrates the difficulty of the dataset, and on the other hand it demonstrates the importance of capturing the spatial relations between body parts, which can improve part localization by more than ten times.

Fig. 6 shows examples of estimated upper-body configurations, which demonstrate the effectiveness of our method even for difficult poses including self-occlusions (*e.g*. examples (a), (c), and (h)). We also show some typical failure cases that are often due to incorrect scale estimation by the weak object detector (example (i)), or failures of the part detector (parts (k) and (l)). Since we assume a constant size of the object parts, our method is limited in how foreshortening can be tolerated (example (j)).

3.3. Full body pose estimation

Finally, we evaluate our model on a full body pose estimation task and compare to the iterative image parsing method of Ramanan [15], which uses a similar spatial model but approaches appearance modeling quite differently. Note that this is the same algorithm that was used in the pose estimation stage of [6]. In this comparison we use the publicly available multi-view and multi-articulation

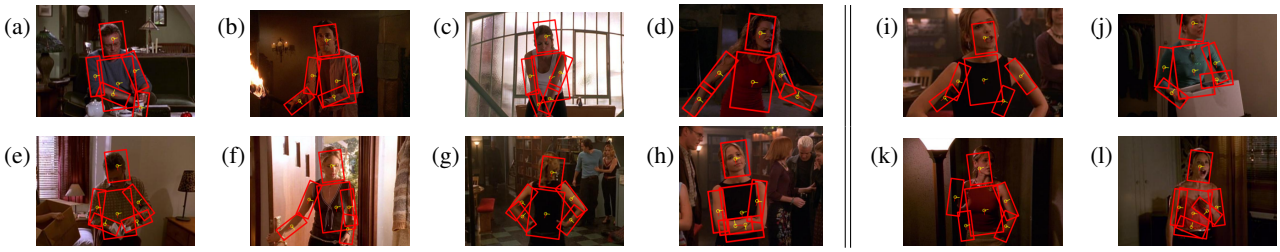


Figure 6. **Upper body pose estimation:** Examples of pose estimation results (left), and some typical failure cases (right) of our method on the "Buffy" dataset (see text for description). Note that color information is not used, but shown here for illustration.

Method	Torso	Upper arm		Forearm		Head	Total
Progressive Search Space Reduction [6]	–	–	–	–	–	–	57.9
Our part detectors	18.9	6.6	7	3.3	2.9	47.2	14.3
Our part detectors and inference with generic prior	**90.7**	80.2	78.4	40.1	42.3	**95.9**	71.3
Our part detectors and inference with front/back view prior	**90.7**	**80.6**	**82.1**	**44.2**	**47.9**	95.5	**73.5**

Table 1. **Upper body pose estimation:** Comparison of body part detection rates on the "Buffy" dataset [6] (numbers indicate the percentage of correctly detected parts. The total number of part segments is $6 \times 269 = 1614$).

Iterative Image Parsing dataset from [15], which contains people engaged in a wide variety of activities ranging from simple standing and walking to dancing and performing acrobatic exercises. The difficulty of the task is further increased by the limited training set of only 100 images, which only scarcely capture the variations in appearance and poses present in the test set. We evaluate the part localization performance using the same criteria as proposed in [6] and used in Sec. 3.2. The iterative image parsing results were obtained using the implementation by the author of [15]. Quantitative results are shown in Tab. 2.

Our findings show significant performance gains: The localization results of our approach surpass those of [15] by more than a factor of 2 (55.2% *vs*. 27.2% accuracy). The localization performance of all body parts is significantly improved, sometimes by a factor of 3. It is interesting to note that our head detector alone (*i.e*., without any kinematic model) has a better localization performance for the head than the full model of [15]. This clearly demonstrates the importance of powerful part detectors for obtaining good overall performance.

Tab. 2 also shows the performance of our method when the boosted part detectors are replaced with discriminatively trained edge templates used in [15]. For this experiment we extracted the responses of the part templates using the author's code [15] and fitted sigmoid functions to the foreground and background responses of each part template in order to make them comparable with one another. The average part localization rate in this experiment is 37.5%[1], which is significantly better than the results of iterative image parsing [15] with the same edge template features. We attribute this to the different representation of the part relationships in our kinematic model. The performance of the full model is still significantly better (55.2%), which again shows that the boosted part detectors contribute substantially to the overall performance.

Fig. 7 shows a comparison between both approaches on an arbitrary set of 12 consecutive test images from the "Iterative Image Parsing" dataset. For each image we also show the posterior distributions for each part and give the number of correctly localized body parts. From these results it appears that Ramanan's method works well in relatively uncluttered images (*e.g*., Fig. 7(i) and (l)); nonetheless, even in those scenes, we often localize more parts correctly. In strongly cluttered scenes (*e.g*., Fig. 7(d) or (g)), our method seems to have a clear advantage in recovering the pose. This may be attributed to the fact that the image parsing approach was not able to build up appropriate appearance models during the initial parse. The line templates used in the initial parse may also be misguided by strong edges (Fig. 7(a)), which our method handles more gracefully.

4. Conclusion

In this paper we proposed a generic model for detection and articulated pose estimation. We demonstrated the generality of our approach on 3 recent datasets, where it outperformed specialized approaches by a large margin, which have been designed specifically for only one of these tasks. Despite that, our model is surprisingly simple. We attribute these excellent results to a powerful combination of two components: A strong discriminatively trained appearance model and a flexible kinematic tree prior on the configurations of body parts. In order to facilitate comparison with our model we will make the source code of our implementation available on our website[2]. Currently, we do not make use of color information and do not model relationships between body parts beyond kinematic constraints, for example

[1] Results without sigmoid fitting are considerably worse.

[2] `www.mis.informatik.tu-darmstadt.de/code`

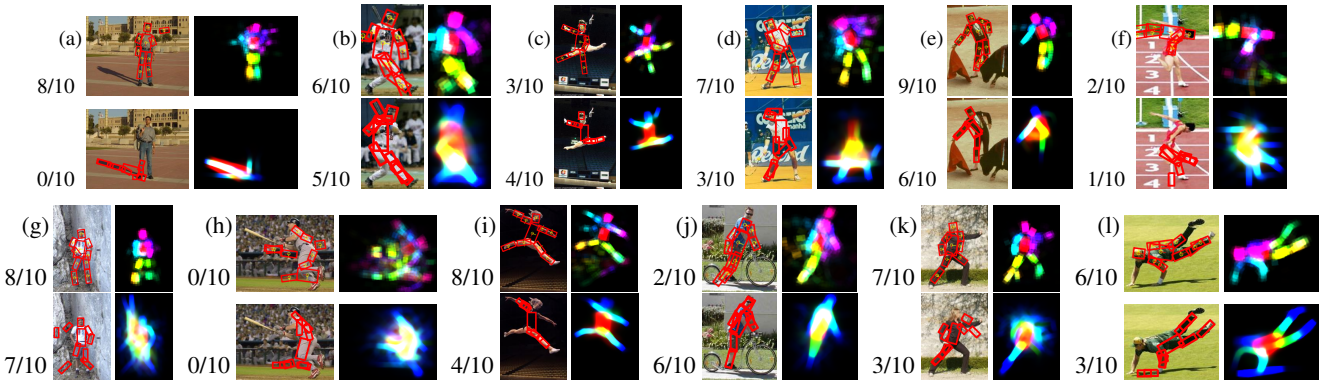


Figure 7. **Comparison of full body pose estimation results** between our approach (top) and [15] (bottom). The numbers on the left of each image indicate the number of correctly localized body parts.

Method	Torso	Upper leg		Lower leg		Upper arm		Forearm		Head	Total
IIP [15], 1st parse (edge features only)	39.5	21.4	20	23.9	17.5	13.6	11.7	12.1	11.2	21.4	19.2
IIP [15], 2nd parse (edge + color feat.)	52.1	30.2	31.7	27.8	30.2	17	18	14.6	12.6	37.5	27.2
Our part detectors	29.7	12.6	12.1	20	17	3.4	3.9	6.3	2.4	40.9	14.8
Our inference, edge features from [15]	63.4	47.3	48.7	41.4	34.14	30.2	23.4	21.4	19.5	45.3	37.5
Our inference, our part detectors	**81.4**	**67.3**	**59**	**63.9**	**46.3**	**47.3**	**47.8**	**31.2**	**32.1**	**75.6**	**55.2**

Table 2. **Full body pose estimation:** Comparison of body part detection rates and evaluation of different components of the model on the "Iterative Image Parsing" (IIP) dataset [15] (numbers indicate the percentage of the correctly detected parts. The total number of part segments is $10 \times 205 = 2050$).

in order to model occlusions (*c.f*. [20]). We expect that such additional constraints will further improve the performance and should be explored in future work.

Acknowledgements: The authors are thankful to Krystian Mikolajczyk for the shape context implementation and Christian Wojek for the AdaBoost code and helpful suggestion. Mykhaylo Andriluka gratefully acknowledges a scholarship from DFG GRK 1362 "Cooperative, Adaptive and Responsive Monitoring in Mixed Mode Environments".

References

[1] M. Andriluka, S. Roth, and B. Schiele. People-tracking-by-detection and people-detection-by-tracking. *CVPR 2008*.

[2] M. Bergtholdt, J. Kappes, S. Schmidt, and C. Schnörr. A study of parts-based object class detection using complete graphs. *IJCV*, 2009. In press.

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

[4] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *IJCV*, 61(1):55–79, 2005.

[5] P. F. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. *CVPR 2008*.

[6] V. Ferrari, M. Marin, and A. Zisserman. Progressive search space reduction for human pose estimation. *CVPR 2008*.

[7] Y. Freund and R. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

[8] H. Jiang and D. R. Martin. Global pose estimation using non-tree models. *CVPR 2008*.

[9] F. R. Kschischang, B. J. Frey, and H.-A. Loelinger. Factor graphs and the sum-product algorithm. *IEEE T. Info. Theory*, 47(2):498–519, Feb. 2001.

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

[11] M. W. Lee and I. Cohen. Proposal maps driven MCMC for estimating human body pose in static images. *CVPR 2004*.

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

[13] K. Mikolajczyk, B. Leibe, and B. Schiele. Multiple object class detection with a generative model. *CVPR 2006*.

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

[15] D. Ramanan. Learning to parse images of articulated objects. *NIPS*2006*.

[16] D. Ramanan and C. Sminchisescu. Training deformable models for localization. *CVPR 2006*.

[17] X. Ren, A. C. Berg, and J. Malik. Recovering human body configurations using pairwise constraints between parts. *ICCV 2005*.

[18] R. Ronfard, C. Schmid, and B. Triggs. Learning to parse pictures of people. *ECCV 2002*.

[19] E. Seemann, B. Leibe, K. Mikolajczyk, and B. Schiele. An evaluation of local shape-based features for pedestrian detection. *BMVC 2005*.

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

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

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

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

[24] P. Viola, M. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. *ICCV 2003*.

[25] J. Zhang, J. Luo, R. Collins, and Y. Liu. Body localization in still images using hierarchical models and hybrid search. *CVPR 2006*.