This CVPR2013 paper is the Open Access version, provided by the Computer Vision Foundation.
The authoritative version of this paper is available in IEEE Xplore.

Human Pose Estimation using Body Parts Dependent Joint Regressors

Matthias Dantone[1] Juergen Gall[2] Christian Leistner[3] Luc Van Gool[1]
ETH Zurich, Switzerland[1] MPI for Intelligent Systems, Germany[2] Microsoft, Austria[3]
{dantone,vangool}@vision.ee.ethz.ch jgall@tue.mpg.de christian.leistner@microsoft.com

Abstract

In this work, we address the problem of estimating 2d human pose from still images. Recent methods that rely on discriminatively trained deformable parts organized in a tree model have shown to be very successful in solving this task. Within such a pictorial structure framework, we address the problem of obtaining good part templates by proposing novel, non-linear joint regressors. In particular, we employ two-layered random forests as joint regressors. The first layer acts as a discriminative, independent body part classifier. The second layer takes the estimated class distributions of the first one into account and is thereby able to predict joint locations by modeling the interdependence and co-occurrence of the parts. This results in a pose estimation framework that takes dependencies between body parts already for joint localization into account and is thus able to circumvent typical ambiguities of tree structures, such as for legs and arms. In the experiments, we demonstrate that our body parts dependent joint regressors achieve a higher joint localization accuracy than tree-based state-of-the-art methods.

1. Introduction

Estimating the human pose from still images is a very active field due to its relevance for applications [21]. One of the most popular approaches in this area is the pictorial structure framework [13, 11], which models the spatial relations of rigid parts using usually a tree model. Pictorial structures have been improved for pose estimation in many ways, e.g., by learning better appearance [24, 9, 1] or shape models [42] of the body parts.

In object detection, one of the best performing methods relies on so called deformable part models [10], which use mixtures of star models over templates of parts. Recently, [40] showed that mixtures of part templates can also be efficiently used in a tree model, leading to very powerful pose estimation models. In particular, instead of modeling the transformations of a single body part template as in the classical pictorial structure model, the transformations of the limbs are encoded by different deformable templates per body part. While this approach outperforms classical pictorial structure models for human pose estimation, it has been shown in [41] that the used templates, which are scanning-window templates trained with linear SVMs on HOG features [7], are very sensitive to noise and limit the performance.

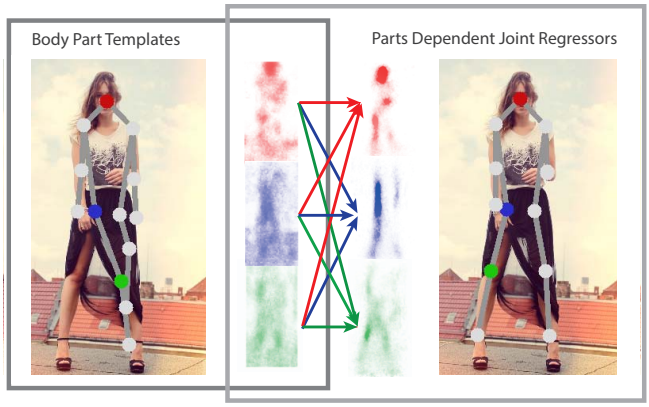

Figure 1. The *dark gray* rectangle on the l.h.s. illustrates a pictorial structure (PS) model with independent part templates. Each classifier estimates independently the probability that an image region belongs to a specific body part, e.g., head (*red*), right hip region (*blue*), and right knee region (*green*). The confidence maps are used as unary potentials for a PS model with 13 joints. Neither the independent classifiers nor the tree structure of the PS model are able to resolve the ambiguities between the left and right leg. The *light gray* rectangle on the r.h.s. illustrates the proposed approach where two layers are used. While the first layer consists of the same independent classifiers, the second layer regresses the locations of the joints in dependency of the independent part classifiers. The confidence maps of the regressed points, e.g., nose (*red*), left hip joint (*blue*), and left knee (*green*), are more discriminative and resolve the ambiguities between the legs.

In this work, we thus address the problem of obtaining better part templates in the context of a pictorial structure framework. Similar to [40], we do not model the limb transformations explicitly, but use discriminative learned templates that allow the handling of limb pose variations im-

plicitly. However, contrary to [40], we do not use noise sensitive, scanning-window templates, but instead propose non-linear regressors for the joint locations. As regressors, we rely on random forests that have shown to be fast, robust, and accurate in the context of predicting body parts or joint locations from depth data [29, 15].

While previous work treats all body part templates independently and uses the pictorial structure framework to model spatial and orientation relations between part templates, we propose a more discriminative template representation that already takes co-occurrences and relations to other parts to some extent into account, as illustrated in Fig. 1. To this end, we train joint regressors that use the output of independent body part templates as input and thus predict the location of a joint in dependency of the co-occurrence of other body parts. In this way, joint regressors are already able to resolve some typical problems of tree models, such as the discrimination of left and right limbs.

In our experiments, we show that the proposed body parts dependent joint regressors achieve a much higher joint localization accuracy than independent part templates or joint regressors. Integrated into a pictorial structure framework, the approach achieves a better joint localization accuracy than a state-of-the-art method [40] at comparable running time of a few seconds per image.

2. Related Work

Human pose estimation is a well studied area with many interesting applications, such as, gaming, human-computer interaction or health care. For a detailed review of various applications and methods, we refer the reader to [21]. In this section, we review only the most related work with a focus on pose estimation within a pictorial structure framework.

Pictorial structure models are well known since the 70s [13] and became very popular with the introduction of efficient inference algorithms [11]. While many approaches relied at the beginning on simple geometric primitives for the body parts and simple color models or background subtraction for the likelihoods, many improvements have been made to the part templates. For instance, linear SVMs for learning discriminative part templates were introduced in [26]. In [18], a cascade of body parts detectors were proposed to obtain more discriminative templates. Other approaches rely on several templates for a single body part [32, 40]. Furthermore, human body models have been used to obtain better shapes of the body parts [42] or to synthesize training data [23]. A variety of image features for pose estimation has been investigated in [1].

Another research direction has focused on introducing richer body models that overcome the limitation of tree structures. For instance, a body part can be assigned with high confidence to two nodes of a tree in case of weak part templates or occlusions, e.g., the left and right body part are sometimes assigned to a single observation. To prevent this, additional constraints between the limbs [31, 25, 17, 34] or even a fully connected graphical model [2, 36] have been proposed. Loopy models, however, make the inference more expensive and require approximations for inference.

Other approaches rely on several models. For instance, several tree models are combined by a boosting procedure in [37], whereas [28] predicts some parameters of the tree model from the image data. The latter approach is related to methods that estimate the pose directly from image features like [3], but also methods that iteratively refine the model by adapting the appearance [24, 9].

Besides of independent part templates for body parts, also hierarchies of part templates have been proposed [33, 38, 35]. [33] also introduces attributes of body parts allowing the sharing of part templates of similar shape. The hierarchy proposed in [38] even discards the semantic meaning of body parts and relies on the concept of poselets [4].

Our work is focused on improving the body part templates or the likelihoods for the joint positions within a pictorial structure model. In contrast to previous works, which run each body part template independently and use a tree structure or loopy models for modeling the dependencies among body parts, we propose to take the dependencies between body parts already into account for predicting the joint locations. In this way, the joint or part templates are already able to discriminate left and right limbs and compensate already for some limitations of tree models. Since the templates are implemented by efficient randomized regression forests that predict directly the joint locations, our approach is comparable in running time to a state-of-the-art method [40], while providing a higher joint localization accuracy.

Random forests have been previously used for pose estimation from depth data [29, 15]. In a similar spirit, an implicit shape model [20] has been used for pose estimation in [22]. Random forests have been also used to improve poselets for pose estimation from depth data [16] and for pedestrian detection [27]. A random forest approach with two layers has been proposed in [30] for image segmentation. While the first layer converts an image into a codeword representation, so-called textons, the second layer performs pixel-wise image segmentation based on the textons.

3. Pictorial Structure

As a human body model, we use a classical pictorial structure framework [11]. However, instead of using a limb representation for the body configuration, we use a joint representation $\mathcal{J} = \{\mathbf{j}_k\}$ where each joint $\mathbf{j}_k = (\mathbf{x}_k)$ encodes the image location of a joint. The root of the tree is defined by the nose, the only non-joint point in the body configuration. The prior on part configurations is therefore

defined by

$$p(\mathcal{J}) = \prod_{(k,l)\in E} \psi_{kl}(\mathbf{j}_k, \mathbf{j}_l), \tag{1}$$

where E are the directed edges of the kinematic chain shown in Fig. 1. As in [11], we model the binary potentials $\psi_{kl}(\mathbf{j}_k, \mathbf{j}_l)$ by Gaussian distributions for efficient inference.

The pose configuration can be estimated from a still image by searching the maximum of the posterior distribution

$$p(\mathcal{J}|\mathbf{I}) \propto p(\mathbf{I}|\mathcal{J})p(\mathcal{J}). \tag{2}$$

Assuming independent part templates for the likelihood, the posterior can be written as

$$p(\mathcal{J}|\mathbf{I}) \propto \prod_{k} \phi_k(\mathbf{j}_k) \cdot \prod_{(k,l)\in E} \psi_{kl}(\mathbf{j}_k, \mathbf{j}_l). \tag{3}$$

The unary potentials $\phi_k(\mathbf{j}_k)$ are in many cases only approximations of the likelihoods $p(\mathbf{I}|\mathbf{j}_k)$ and correspond to part templates. For instance, HOG features [7] and linear SVMs are used as part templates in [40]. While we use Gaussian binary potentials and perform inference as in [10], our work focuses only on extracting more discriminative unary potentials $\phi_k(\mathbf{j}_k)$. In particular, we address the weakness of independent part templates and propose non-linear, parts dependent joint regressors instead.

4. Joint Regressors

A joint representation as in (1) has the advantage that limb transformations like foreshortening do not need to be explicitly modeled in the pictorial structure model, which reduces complexity and running time. The independence assumption of common part templates is relaxed by training the regressors on image features and confidence maps of other body parts, i.e.,

$$\phi_k(\mathbf{j}_k) = p(\mathbf{j}_k|\mathbf{I}, \mathcal{L}), \tag{4}$$

where $\mathcal{L}$ is the set of body parts. In this work, we use the term 'joint' for any landmark point like a skeleton joint or the nose, whereas 'body parts' are defined as regions around the joints as illustrated Fig. 1.

As regressors, we use random forests [5]. For completeness, we give a brief introduction to random forests in Section 4.1. In Sections 4.2, 4.3, and 4.4, we discuss three variations, namely part templates using random forests, independent joint regressors, and parts dependent joint regressors.

4.1. Random Forests

Random forests [5] or in general decision forests [6] have been used for many classification or regression tasks, for instance, labeling body parts in depth images [29], predicting the joint positions from depth data [15], or localizing facial feature points [8]. In this section, we describe the general training procedure and discuss the details regarding used features, split functions, etc. in the following sections.

Random forests are ensembles of randomized decision trees that learn a mapping from an image patch P to a distribution over a parameter space Θ. For classifying body parts, the parameter space is the set of class labels or body parts. For predicting the location of a single joint, the parameter space is $\mathbb{R}^2$. To learn such a mapping, a tree T in a forest $\mathcal{T}$ is built from a set of image patches $\mathcal{P}$ that are extracted randomly from a random subset of the training images. Each patch contains a set of image features F_P, such as HOG or color information, and the parameters $\theta_P \in \Theta$ to estimate. During the training of the tree, a set of patches is divided recursively into two subset $\mathcal{P}_L$ and $\mathcal{P}_R$ using a binary split function $\zeta^*(F_P) \rightarrow \{0, 1\}$, which is defined on the patch features. Every split function is chosen from a randomly generated set of split functions $\{\zeta\}$ by maximizing the goodness or information gain of the split $g(\zeta)$:

$$\zeta^* = \arg\max_{\zeta} g\left(\zeta\right), \tag{5}$$

$$g\left(\zeta\right) = \mathcal{H}\left(\mathcal{P}\right) - \sum_{S\in\{L,R\}} \frac{|\mathcal{P}_S\left(\zeta\right)|}{|\mathcal{P}|} \mathcal{H}\left(\mathcal{P}_S\left(\zeta\right)\right), \tag{6}$$

where $\mathcal{H}$ is, depending on Θ, the entropy or the sum-of-squared-differences. After the split, the binary function is stored at the node and the training continues recursively until the maximum depth of the tree is reached or the gain drops below a predefined threshold. At the leaves, the distributions $p(\theta|L)$ are estimated based on the parameters of the patches $\mathcal{P}$ arriving at the leaf L.

4.2. Body Part Templates

The body part templates are modeled as classical limb templates trained with a random forest. As patch feature, we use a set of features $F_P = F_P^f$ that is inspired by [14], where F_P^f is a matrix of fixed size containing the values of the feature f. We use overall 17 features: a normalized gray-scale version of the image; the Lab color space where each color channel is processed by a min and a max filtration with 5x5 filter size; HOG with 9 bins using a 5x5 cell and soft binning. The values of each bin of HOG are mapped to a matrix F_P^f and processed by a max filter. Additionally to the color and HOG features, we added the output of a skin detector [19] as feature. We train a separate forest for each body part, where each forest is trained by body part patches sampled from a Gaussian distribution centered at the body part annotation and negative patches sampled uniformly from the background of the image. Each patch P is therefore augmented by a binary label c, which is k if it is sampled from body part $\mathbf{l}_k$. We use the same number of body parts as joints, i.e., 13.

The used split functions are pixel comparisons as in [14]:

$$\zeta_\gamma(P) = \begin{cases} 1 & F_P^f(\boldsymbol{q}) - F_P^f(\boldsymbol{p}) < \tau \\ 0 & \text{otherwise,} \end{cases} \tag{7}$$

where the parameters $\gamma = (\boldsymbol{p}, \boldsymbol{q}, f, \tau)$ describe two coordinates $\boldsymbol{p}$ and $\boldsymbol{q}$ within the patch boundaries, the selected appearance channel $f \in \{1, 2, \ldots, C\}$, and the defined threshold τ, respectively. For selecting the binary tests (6), we use the entropy

$$\mathcal{H}(\mathcal{P}) = -\sum_c p(c|\mathcal{P}) \log\left(p(c|\mathcal{P})\right). \tag{8}$$

The unary potentials for the body parts $\mathbf{l}_k$ are obtained by densely extracting image patches from the test image and passing them through the trained trees. A single patch P ends at a leaf L_T for each tree T. Based on the class probabilities $p(c|L_T)$ stored at the leaves, the unary potential at patch location $\mathbf{x}_P$ is defined by the average probability of all trees in the forest:

$$\phi_k(\mathbf{l}_k(\mathbf{x}_P)) = \frac{1}{|\mathcal{T}|} \sum_{T \in \mathcal{T}} p(c{=}k|L_T(P)). \tag{9}$$

Averaging the class probabilities of the trees is a common approach for random forests [5].

4.3. Independent Joint Regressors

For the regression, a sampled patch P is additionally augmented with an offset vector $\mathbf{v}_{P,k}$ pointing to the location of the corresponding joint $\mathbf{j}_k$. During training, the goodness (6) for evaluating the split functions is based on the sum-of-squared-distances; that is

$$\mathcal{H}(\mathcal{P}) = \frac{1}{|\mathcal{P}|} \sum_{P \in \mathcal{P}} \|\mathbf{v}_{P,k} - \mu_k\|^2, \tag{10}$$

where μ_k denotes the mean. At the leaves, the class probabilities $p(c|L_T)$ and the probabilities over the offset vectors $p(\mathbf{v}|L_T)$ are stored. The unary potential at location $\mathbf{x}$ for joint k is defined by

$$\phi_k(\mathbf{j}_k(\mathbf{x})) = \sum_{\mathbf{y} \in \Omega} \frac{1}{|\mathcal{T}|} \sum_{T \in \mathcal{T}} \Big\{ p\left(c{=}k|L_T(P(\mathbf{y}))\right) \cdot p\left(\mathbf{x}{-}\mathbf{y}|L_T(P(\mathbf{y}))\right) \Big\}. \tag{11}$$

After computing the unary potentials for an image, the unary potentials for each joint are normalized to be within the range $[0, 1]$. During training, a random forest can minimize both splitting criteria, *i.e.*, (8) and (10), simultaneously. This is achieved simply via randomly alternating between the two goodness measures while the samples are recursively split down the tree, *c.f.* [14].

Figure 2. Sample images from the FashionPose dataset with annotations. The *red* circles bottom right show the error thresholds 0.1, 0.15, and 0.25 used for evaluation.

4.4. Parts Dependent Joint Regressors

The previous part potentials are calculated independently. That is, during both training and evaluation, each sampled patch is evaluated without taking its spatially surrounding potentials into account. For the task of joint localization, this can result in ambiguities, *e.g.*, for left and right knees as illustrated in Fig. 1. To resolve this issue, we propose a third potential that predicts the joint locations as in (11), but also takes neighboring part potentials into account:

$$\phi_k(\mathbf{j}_k, \mathcal{L}) = p(\mathbf{j}_k|\mathbf{I}, \mathcal{L}) \tag{12}$$

However, incorporating a multi-dimensional neighborhood structure is usually computationally demanding. Therefore, we approximate (12) by splitting our regression model into two layers. The first layer only calculates independent part potentials $\phi_k(\mathbf{l}_k)$ (9). The second layer also predicts unary potentials but also incorporates the potentials of the first layer and their locations as additional feature maps. Thus the set of training patches for the second forest can be written as $\{P = (\mathcal{F}_P^*, c_P, \mathbf{v}_P)\}$, where $\mathcal{F}_P^* = \{1, \ldots, C; \Phi_1, \ldots, \Phi_k\}$ is the enriched set of feature channels. The leaf probabilities $p(c|\mathcal{L}, L_T)$ and $p(\mathbf{v}|\mathcal{L}, L_T)$ now depend on the probabilities of the body parts and we obtain

$$\phi_k(\mathbf{j}_k, \mathcal{L}) = \sum_{\mathbf{y} \in \Omega} \frac{1}{|\mathcal{T}|} \sum_{T \in \mathcal{T}} \Big\{ p\left(c{=}k|\mathcal{L}, L_T(P(\mathbf{y}))\right) \cdot p\left(\mathbf{x}{-}\mathbf{y}|\mathcal{L}, L_T(P(\mathbf{y}))\right) \Big\}. \tag{13}$$

5. Experiments

For evaluation we use two datasets, namely the well-known Leeds Sports Pose dataset (LSP [18]) and a newly

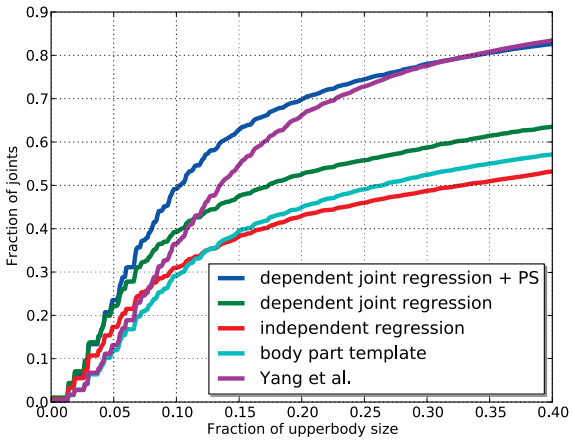

Figure 3. Comparison of the joint localization accuracy of the proposed unary potentials and comparison with a state-of-the-art method [40]. While the body part classification (9) and the independent joint regression (11) perform similarly, they are drastically outperformed by the proposed body parts dependent joint regressors (13). Since the body parts dependent joint regressors do not encode any explicit information of the human skeleton, using a pictorial structure model (PS), which models the kinematic chain, gives an additional performance boost. The body parts dependent joint regression together with a pictorial structure model outperforms [40]. In particular at low error rates like 0.1, the number of correctly localized joints is 20% higher than [40].

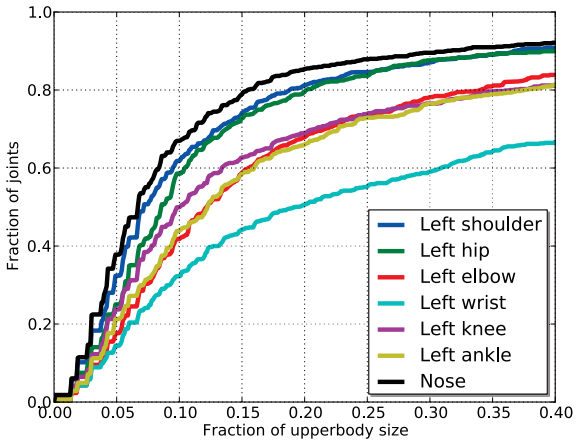

Figure 4. The accuracy plots for individual joints using body parts dependent joint regressors with a pictorial structure model. For better readability, we plot only the left joints. As expected, localizing the wrist is the most difficult task, whereas head, shoulders, and hip joints are reasonable well localized. The numbers for all joints at error thresholds 0.1 and 0.15 are provided in Table 1.

collected dataset that we call *FashionPose*. While the LSP dataset contains a high variation of poses, the variation of appearance and dress style within each of the eight sport classes is rather small. We have therefore collected a new dataset that has very high variation in cloth and appearance. In our experiments, we compare our method to three related methods, namely linear and non-linear SVMs for part templates [18] and flexible mixtures-of-parts [40]. We also compare our approach to two other state-of-the-art methods, namely pose-specific part appearance classifiers [18] and spatial hierarchies of mixture models [35].

FashionPose Dataset. Since clothing imposes a particular challenge for pose estimation in general, which is not well reflected in current datasets for pose estimation from still images, we collected a new dataset. The proposed dataset consists of 7,543 accurate annotated images downloaded from a variety of fashion blogs, e.g., `lookbook.nu` and `kalei.do`. Each image contains a person where the full body is visible and is annotated by 12 joints and a point for the head, namely the nose. We did not annotate the head by the top of the head and the neck as in other datasets [39, 18] since these two points were very difficult to annotate accurately. Occluded joints have also been annotated.

The dataset is not only challenging due to the large variation of dressing style ranging from casual dresses and gowns to haute couture, but it also contains a large variation of poses. For evaluation, we grouped the dataset into a training set containing 6,543 images and a set of 1,000 testing images and rescaled all images to a common upper body size of 75 pixels, measured by the distance between the average position of the two hip joints and the average position of the two shoulder joints. The dataset is more challenging than the Fashionista dataset [39] that contains only 685 images. While the Fashionista dataset has been proposed for parsing clothes and not for pose estimation, the FashionPose dataset can be also augmented with additional annotations for evaluating methods for parsing clothes in still images as well. The FashionPose dataset is publicly available.[1] Some example images with ground truth annotation are shown in Fig. 2.

Evaluation measurement. In our experiments, we measure the joint localization error as a fraction of the upper body size. This measurement is well established for other computer vision tasks, e.g., fiducial point detection. It is independent of the actual size of the image and more precise than common measures derived from bounding box-based object detection like PCP [12]. PCP declares a limb as correctly detected if the error of the predicted endpoints are within 50% of the limb length from the ground truth endpoints. We use the imprecise PCP measure only for comparison with other reported results on the Leeds Sports Pose dataset; otherwise we use the more informative normalized joint localization error.

Experiments on FashionPose. For the training of the

[1] `http://www.vision.ee.ethz.ch/~mdantone/fashionpose`

error thres. joints	0.10 ours	0.10 Yang et al.	0.15 ours	0.15 Yang et al.
Head	**66.97**	56.16	**78.84**	77.76
L. shoulder	**61.94**	53.21	**73.81**	72.75
R. shoulder	**61.81**	55.39	**74.19**	74.03
Left hip	**57.16**	38.43	**72.90**	58.61
Right hip	**58.58**	34.96	**73.81**	58.09
Left elbow	**41.81**	27.89	**56.00**	46.14
Right elbow	**41.29**	32.51	**58.84**	50.64
Left wrist	**32.26**	24.29	**44.26**	38.17
Right wrist	**29.68**	21.72	**39.48**	33.16
Left knee	**52.13**	39.07	**65.29**	56.94
Right knee	**49.94**	38.43	**62.71**	57.32
Left ankle	**43.87**	32.26	**58.97**	49.61
Right ankle	**41.68**	31.10	**58.58**	48.20

Table 1. Detection accuracy for all joints at error thresholds 0.1 and 0.15. The comparison shows that our method performs similar or better than [40] for all joints.

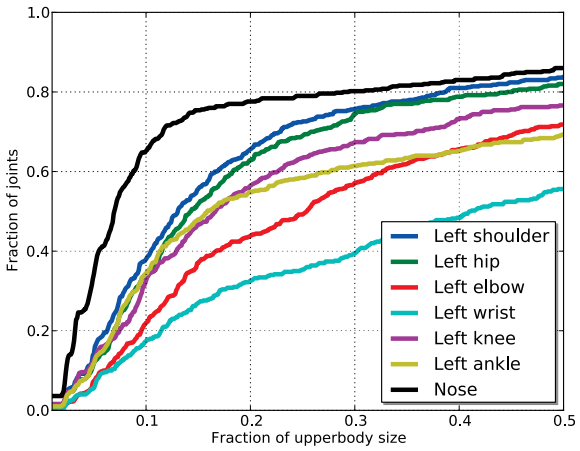

Figure 5. Accuracy plots for individual joints on the LSP dataset.

random forests for the body part templates, independent and parts dependent joint regression, we fixed some parameters intuitively. The patch size, and thus of the feature matrices F_P^f, is 30x30 pixels. Each forest consists of 15 trees with maximum depth of 20 and a minimum number of 20 patches per leaf. For training, we generate 25,000 binary tests (7) at each node, where we use 1,000 random parameter settings for $\gamma \backslash \tau$ and for each setting additionally 25 random thresholds τ. Each tree has been grown on a set of 500,000 positive and 500,000 negative patches extracted from 4,000 randomly selected training images. For computational reasons, we evaluate the split functions at each node for only maximal 200,000 patches.

We first evaluated the performance of the part templates (Section 4.2), the independent joint regressors (Section 4.3), and the body parts dependent joint regressors (Section 4.4). The accuracy based on the normalized joint estimation error is given in Fig. 3. The proposed body parts dependent joint regressors clearly outperform the independent part templates and joint regressors. Integrating them into a pictorial structure model (Section 3), which encodes the kinematic skeleton, improves the accuracy further. The accuracy curves for individual joints are plotted in Fig. 4. We also evaluated the accuracy when the unary potentials for classification (9) and independent regression (11) are multiplied. In this case, the performance has not improved compared to the individual unary potentials. This shows that training the regressors depending on the body part templates (13) is essential for the performance gain.

We also compared our approach to a state-of-the-art method proposed by Yang et al. [40] that uses a flexible mixture of templates modeled by linear SVMs. For a fair comparison, we trained the publicly available source code on the entire 6,543 rescaled training images. A comparison of the approach [40] and the parts dependent joint regression is shown in Fig. 3. For an error threshold up to 0.25, the pictorial structure model with parts dependent joint regression outperforms [40]. Larger error thresholds indicate a poor accuracy that is probably insufficient for applications; see Fig. 2. For error thresholds like 0.1, the accuracy is improved by more than 20%. Table 1 compares the accuracies for all joints at error thresholds 0.1 and 0.15. Our approach localizes the joints with a higher accuracy. We also experimented with a three-layer system, but we could not see a significant improvement (+0.3%).

Experiments on Leeds Sports Pose Dataset. Due to the small size of the LSP dataset [18], we trained only 10 trees using 100,000 positive and 100,000 negative patches sampled from the 1,000 training images. The other parameters are the same used for the FashionPose dataset. In order to compare with previous works, we use the PCP criteria. To this end, we added the neck and the top of the head as joints and converted our joint representation into a limb representation by using the joints as endpoints of the limbs. The torso is obtained by the line between the average position of the two hip joints and the average position of the two shoulder joints.

The results of our method using body parts dependent joint regression with a pictorial structure are given in Table 2. The comparison with a pictorial structure model that uses linear SVMs [18] or a cascade of non-linear SVMs [18] as part templates shows that our proposed unary potentials achieve a much higher accuracy. The accuracy with respect to the normalized joint localization error for individual joints is plotted in Fig. 5.

We also compare our approach with the state-of-the-art on this dataset. In [18], the pose data has been also clustered to train a model for each cluster. As can be seen from Table 2, this increases the performance by around 20%. The performance gain can be also explained by the dataset that contains eight different sports classes that are very distinct in appearance and poses. Nevertheless, our approach

Limps	Avg.	Torso	Upper Leg		Lower Leg		Upper Arm		Forearm		Head
Related methods											
Linear SVM [18]	36.4	64.1	42.4	43.1	41.2	40.7	26.2	23.7	16.5	15.7	49.9
Non-linear SVM [18]	44.7	70.9	53.5	58.7	49.3	47.4	37.1	29.1	**26.8**	18.8	55.9
Proposed	**55.5**	**81.6**	**66.0**	**67.0**	**60.8**	**61.2**	**46.4**	**43.8**	25.6	**23.8**	**79.2**
State-of-the-art											
Cluster + Linear SVM [18]	43.6	74.1	54.4	53.6	49.0	49.3	30.5	30.9	17.5	17.7	59.7
Cluster + Non-linear SVM [18]	55.1	78.1	64.8	66.7	60.3	57.3	48.3	46.5	34.5	31.2	62.9
Spatial Hierarchy of Mixture Models [35]	58.8	93.7	68.0		57.8		49.0		29.2		86.5
Cluster + S. Hierarchy of M. M. [35]	61.3	95.8	69.9		60.0		51.9		32.9		87.8

Table 2. Detection accuracy on the Leeds Sports Pose dataset. For comparison, we converted our estimated joint positions into a limb representation and use PCP as measure. For more details regarding the evaluation, we refer to the text. Our method outperforms related methods using linear or non-linear SVMs for part templates within a pictorial structure framework. Only [35] achieves a better performance, but this approach uses a more complex and more expensive model than pictorial structures with a tree structure.

Figure 6. Qualitative results on some representative images from the *FashionPose* and the LSP dataset.

already achieves comparable results with a single model. [35] uses a more complex model than a tree structure that captures the space of plausible human poses much better. While this method achieves better results on this dataset, this comes probably at the cost of higher training and running times. Since the focus of this work is the improvement of the unary potentials in a pictorial structure framework, we used only a single tree model and have not performed clustering or used a more complex body model. However, we expect that also more complex models benefit from better part or joint templates.

6. Conclusion

In this paper, we have addressed robust human pose estimation from still images by proposing novel discriminative part template predictors within a pictorial structure framework. Our joint location regressors consist of random forests that operate over two layers. While the first layer acts as an independent body part classificator, the second one takes the predicted distributions of the first layer for estimating the joint locations into account, thus allowing to put the body parts into relation. In the experimental part, we have shown that our model yields higher accurate human joint predictors than independent part templates and outperforms state-of-the-art methods that also use a tree structure for the human model.

Acknowledgements The authors acknowledge financial support from the ERC Grant (VarCity), the EC projects RADHAR (FP7-ICT-248873) and TANGO TANGO (FP7-ICT-249858) and the CTI project (12618.1 PFES-ES).

References

[1] M. Andriluka, S. Roth, and B. Schiele. Discriminative appearance models for pictorial structures. *IJCV*, 99(3):259–280, 2012. 1, 2

[2] M. Bergtholdt, J. H. Kappes, S. Schmidt, and C. Schnörr. A study of parts-based object class detection using complete graphs. *IJCV*, 87(1-2):93–117, 2010. 2

[3] L. Bo and C. Sminchisescu. Twin gaussian processes for structured prediction. *IJCV*, 87(1-2):28–52, 2010. 2

[4] L. Bourdev, S. Maji, T. Brox, and J. Malik. Detecting people using mutually consistent poselet activations. In *ECCV*, pages 168–181, 2010. 2

[5] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. 3, 4

[6] A. Criminisi, J. Shotton, and E. Konukoglu. Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. *Found. Trends. Comput. Graph. Vis.*, 7(2–3):81–227, 2012. 3

[7] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, pages 886–893, 2005. 1, 3

[8] M. Dantone, J. Gall, G. Fanelli, and L. Van Gool. Real-time facial feature detection using conditional regression forests. In *CVPR*, pages 2578–2585, 2012. 3

[9] M. Eichner and V. Ferrari. Better appearance models for pictorial structures. In *BMVC*, 2009. 1, 2

[10] P. F. Felzenszwalb, R. B. Girshick, D. A. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *TPAMI*, 32(9):1627–1645, 2010. 1, 3

[11] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *IJCV*, 61(1):55–79, 2005. 1, 2, 3

[12] V. Ferrari, M. Marin-Jimenez, and A. Zisserman. Progressive search space reduction for human pose estimation. In *CVPR*, 2008. 5

[13] M. A. Fischler and R. A. Elschlager. The representation and matching of pictorial structures. *IEEE Trans. Comput.*, 22(1):67–92, 1973. 1, 2

[14] J. Gall, A. Yao, N. Razavi, L. Van Gool, and V. Lempitsky. Hough forests for object detection, tracking, and action recognition. *TPAMI*, 33(11):2188–2202, 2011. 3, 4

[15] R. Girshick, J. Shotton, P. Kohli, A. Criminisi, and A. Fitzgibbon. Efficient regression of general-activity human poses from depth images. *ICCV*, pages 415–422, 2011. 2, 3

[16] B. Holt, E.-J. Ong, H. Cooper, and R. Bowden. Putting the pieces together: Connected poselets for human pose estimation. In *Workshop on Consumer Depth Cameras for Computer Vision*, pages 1196–1201, 2011. 2

[17] H. Jiang and D. Martin. Global pose estimation using non-tree models. In *CVPR*, 2008. 2

[18] S. Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *BMVC*, 2010. 2, 4, 5, 6, 7

[19] M. Jones and J. Rehg. Statistical color models with application to skin detection. *IJCV*, 46(1):81–96, 2002. 3

[20] B. Leibe, A. Leonardis, and B. Schiele. Robust object detection with interleaved categorization and segmentation. *IJCV*, 77(1-3):259–289, 2008. 2

[21] T. B. Moeslund, A. Hilton, V. Krüger, and L. Sigal, editors. *Visual Analysis of Humans - Looking at People*. Springer, 2011. 1, 2

[22] J. Müller and M. Arens. Human pose estimation with implicit shape models. In *Workshop on Analysis and Retrieval of Tracked Events and Motion in Imagery Streams*, pages 9–14, 2010. 2

[23] L. Pishchulin, A. Jain, M. Andriluka, T. Thormählen, and B. Schiele. Articulated people detection and pose estimation: Reshaping the future. In *CVPR*, pages 3178–3185, 2012. 2

[24] D. Ramanan. Learning to parse images of articulated bodies. In *NIPS*, pages 1129–1136, 2006. 1, 2

[25] X. Ren, A. C. Berg, and J. Malik. Recovering human body configurations using pairwise constraints between parts. In *ICCV*, pages 824–831, 2005. 2

[26] R. Ronfard, C. Schmid, and B. Triggs. Learning to parse pictures of people. In *ECCV*, pages 700–714, 2002. 2

[27] E. Sangineto, M. Cristani, A. Del Bue, and V. Murino. Learning discriminative spatial relations for detector dictionaries: an application to pedestrian detection. In *ECCV*, pages 273–286, 2012. 2

[28] B. Sapp, C. Jordan, and B. Taskar. Adaptive pose priors for pictorial structures. In *CVPR*, pages 422–429, 2010. 2

[29] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. In *CVPR*, pages 1297–1304, 2011. 2, 3

[30] J. Shotton, M. Johnson, and R. Cipolla. Semantic texton forests for image categorization and segmentation. In *CVPR*, 2008. 2

[31] L. Sigal and M. Black. Measure locally, reason globally: Occlusion-sensitive articulated pose estimation. In *CVPR*, pages 2041–2048, 2006. 2

[32] V. K. Singh, R. Nevatia, and C. Huang. Efficient inference with multiple heterogeneous part detectors for human pose estimation. In *ECCV*, pages 314–327, 2010. 2

[33] M. Sun and S. Savarese. Articulated part-based model for joint object detection and pose estimation. In *ICCV*, pages 723–730, 2011. 2

[34] T.-P. Tian and S. Sclaroff. Fast globally optimal 2d human detection with loopy graph models. In *CVPR*, pages 81–88, 2010. 2

[35] Y. Tian, C. L. Zitnick, and S. G. Narasimhan. Exploring the spatial hierarchy of mixture models for human pose estimation. In *ECCV*, pages 256–269, 2012. 2, 5, 7

[36] D. Tran and D. Forsyth. Improved human parsing with a full relational model. In *ECCV*, pages 227–240, 2010. 2

[37] Y. Wang and G. Mori. Multiple tree models for occlusion and spatial constraints in human pose estimation. In *ECCV*, pages 710–724, 2008. 2

[38] Y. Wang, D. Tran, and Z. Liao. Learning hierarchical poselets for human parsing. In *CVPR*, pages 1705–1712, 2011. 2

[39] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, and T. L. Berg. Parsing clothing in fashion photographs. In *CVPR*, pages 3570–3577, 2012. 5

[40] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *CVPR*, pages 1385–1392, 2011. 1, 2, 3, 5, 6

[41] X. Zhu, C. Vondrick, D. Ramanan, and C. C. Fowlkes. Do we need more training data or better models for object detection? In *BMVC*, 2012. 1

[42] S. Zuffi, O. Freifeld, and M. J. Black. From pictorial structures to deformable structures. In *CVPR*, pages 3546–3553, 2012. 1, 2