Body Parts Dependent Joint Regressors for Human Pose Estimation in Still Images

Body Parts Dependent Joint Regressors for Human Pose Estimation in Still Images

Matthias Dantone, Juergen Gall, *Member, IEEE* Christian Leistner, and Luc Van Gool, *Member, IEEE*

Abstract—In this work, we address the problem of estimating 2d human pose from still images. Articulated body pose estimation is challenging due to the large variation in body poses and appearances of the different body parts. Recent methods that rely on the pictorial structure framework have shown to be very successful in solving this task. They model the body part appearances using discriminatively trained, independent part templates and the spatial relations of the body parts using a tree model. Within such a framework, we address the problem of obtaining better part templates which are able to handle a very high variation in appearance. To this end, we introduce parts dependent body joint regressors which are random forests that operate over two layers. While the first layer acts as an 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 helps to overcome typical ambiguities of tree structures, such as self-similarities of legs and arms. In addition, we introduce a novel dataset termed $FashionPose$ that contains over $7,000$ images with a challenging variation of body part appearances due to a large variation of dressing styles. In the experiments, we demonstrate that the proposed parts dependent joint regressors outperform independent classifiers or regressors. The method also performs better or similar to the state-of-the-art in terms of accuracy, while running with a couple of frames per second.

Index Terms—Human pose estimation, fashion, random forest, regression, classification

1 Introduction

While current systems for human pose estimation achieve impressive results on depth data [1] or multi-camera video footage [2], human pose estimation from still images is still an unsolved task. In particular, images from the web impose many difficulties due to large variation of poses and dressing styles.

In order to address the problem, higher knowledge of the human body is often exploited by modeling humans by body parts that are connected via a skeleton structure. One of the most popular and influential model for human bodies is the pictorial structure model [3], [4] that models the spatial relations of rigid body parts using a tree model. While in [4] each limb is represented by a single template that is parameterized by location, orientation, shape parameters, and an appearance model, Yang and Ramanan [5] propose mixtures of part templates where each body part is represented by a set of deformable part templates. Although this approach performs very well in comparison to classical pictorial structure models for human pose estimation, it has some limitations. For instance, the used scanning-window templates trained with linear SVMs on HOG features [6] are very sensitive to noise [7].

- *Matthias Dantone and Luc V. Gool are with the Computer Vision Laboratory, ETH Zurich, Switzerland.*
  E-mail: {dantone, vangool}@vision.ee.ethz.ch
- *Juergen Gall is with the Computer Vision Group, University of Bonn, Germany.*
  E-mail: gall@iai.uni-bonn.de
- *Christian Leistner is with Microsoft Austria.*

While having many templates per body part compensates partially for it, we propose non-linear regressors for the joint locations instead of many linear part templates. As regressors, we rely on random forests that have shown to be fast, robust to noise, and accurate in the context of predicting body parts or joint locations from depth data [8], [9]. The particular choice of regressors enables us to train our model also on large datasets in reasonable time, which is an important feature for real-world applications and a limitation of many other approaches.

Non-linear regressors, however, can not resolve some ambiguities that are present in human pose estimation by themselves. For instance, a body part can be assigned with high confidence to two nodes of the pictorial structure model in case of weak part templates or occlusions, e.g., as illustrated in Figure 1, the left and right body part are sometimes assigned to a single observation. This has been addressed by proposing richer structures that aim to resolve the weakness of a tree model by adding additional constraints between the limbs [10], [11], [12], [13] or using a fully connected graphical model [14], [15]. Loopy models, however, make the inference more expensive and require approximations for inference. Instead of treating all body part templates independently and modeling the spatial and orientation relations between part templates by a loopy model, we therefore propose a more discriminative template representation that already takes co-occurrences and relations to other parts to some extent into account, as illustrated in Figure 1. To this end, we train joint regressors that use the output of independent body part templates as input and thus predict the

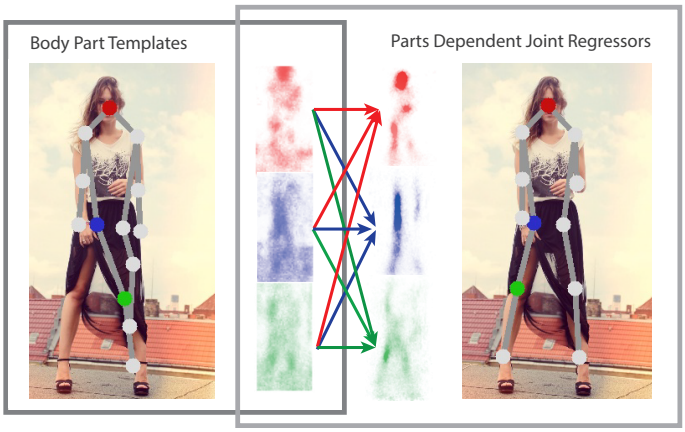


Fig. 1. The *dark gray* rectangle on the l.h.s. illustrates a pictorial structure (PS) model with independent part templates. Each classifier estimates independently the probability that an image region belongs to a specific body part, e.g., head (*red*), right hip region (*blue*), and right knee region (*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.

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.

A preliminary version of this paper appeared in [16]. The present work improves the performance of the approach [16] by using a mixture of tree models for the pictorial structure framework similar to [17]. The runtime and accuracy performance of our approach is thoroughly evaluated on two datasets, namely the well-known Leeds Sports Pose dataset [17] and our newly collected *FashionPose* dataset that captures a large variation of dressing style ranging from casual dresses and gowns to haute couture. The comparison with other methods reveals that our approach achieves a body part detection accuracy on a par with the state-of-the-art on the sports dataset and greatly outperforms [5] in terms of joint localization accuracy on the *FashionPose* dataset, while achieving a run time of a couple of images per second on the LSP dataset.

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 [18]. 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 [3] and became very popular with the introduction of efficient inference algorithms [4]. 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 [19]. In [17], a cascade of body parts detectors were proposed to obtain more discriminative templates. Other approaches rely on several templates for a single body part [5], [20]. In [21] the method of [5] is extended by improving the hard negative mining and exploiting appearance similarities of background and foreground across multiple images.

Furthermore, human body models have been used to obtain better shapes of the body parts [22] or to synthesize training data [23]. A pictorial structure framework for 2d pose estimation is extended to 3d in [24]. A variety of image features for pose estimation has been investigated in [25].

Another research direction has focused on introducing richer body models that overcome the limitation of tree structures. For instance, additional constraints between the limbs [10], [11], [12], [13] or even a fully connected graphical model [14], [15] have been proposed. These 'loopy' models, however, make the inference more expensive and often require approximations for inference. In [26] an efficient and exact inference algorithm based on branch-and-bound has been used to solve the inference in loopy graphical models. In [27] a fully non-parametric Bayesian has been used to model a prior of human pose.

Other approaches rely on model combination. For instance, several tree models are combined by a boosting procedure in [28] and latent tree models are learned to approximate the joint distribution of body part locations in [29]. [30] 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 [31], but also methods that iteratively refine the model by adapting the appearance [32], [33].

Besides of independent part templates for body parts, also hierarchies of part templates have been proposed [34], [35], [36]. [34] also introduces attributes of body parts allowing the sharing of part templates of similar shape. The hierarchy proposed in [35] even

discards the semantic meaning of body parts and relies on the concept of poselets [37]. In [38], [39] an existing pictorial structure is improved by using a poselets representation and in [40] poselets have been used to predict the pose of both arms. The recent work [41] introduces the concept of 'visual symbols' that facilitates geometric context modeling.

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, e.g., 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 joint localization accuracy to several state-of-the-art methods, while achieving a running time at a few frames per second.

Random forests [42] have been previously used for body pose estimation in [8], [9], [43], [44]. The authors of [8], [9] describe a system for real time pose estimation from depth data. In [43] a hierarchical tree is used for exemplar-based human pose estimation in 2d still images and in [44] the same authors jointly localize and recognize the pose of humans using randomized hierarchical cascades classifier that randomly select part-based weak classifiers in a hierarchical way to return a distribution over human poses. In a similar spirit, an implicit shape model [45] has been used for pose estimation in [46].

Random forests have been also used to improve poselets for pose estimation from depth data [47] and for pedestrian detection [48], [49]. A random forest approach with two layers has been proposed in [50] for hand pose classification and estimation and in [51] for image segmentation. In [51] the first layer converts an image into a codeword representation, so-called textons, and the second layer performs pixel-wise image segmentation based on the textons.

Regression forests have recently been used for a variety of applications including real-time face analysis from depth data [52] and 2d images [53], model fitting [54], multi-object segmentation [55], object detection [56], and articulated hand pose estimation [57].

3 Pictorial Structure

As a human body model, we use a classical pictorial structure framework [4]. 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 k. 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 Figure 1. As in [4], 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 $\mathbf{I}$ 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 [6] and linear SVMs are used as part templates in [5]. Section 4 focuses 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. The binary potentials $\psi_{kl}(\mathbf{j}_k, \mathbf{j}_l)$ are discussed in Section 5.

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.,

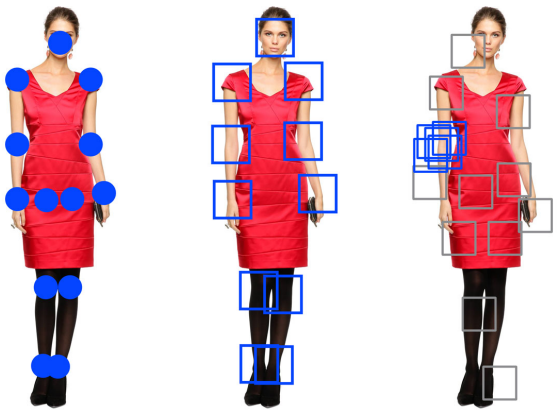
$$\phi_k(\mathbf{j}_k) = p(\mathbf{j}_k|\mathbf{I}, \mathcal{L}), \tag{4}$$

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 Figure 2.

As regressors, we use random forests [42], [58]. 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 [42], [58] or in general decision forests [59] have been used for many classification or regression tasks, for instance, labeling body parts in depth images [8], predicting the joint positions from depth data [9], or localizing facial feature points [53]. In this section, we describe the general training procedure



(a) (b) (c)

Fig. 2. In this work, the term 'joint' refers to a landmark point (a) and 'body part' refers to a region around a landmark (b). (c) During training, positive (blue) patches are sampled around a landmark and negative (gray) patches from the background.

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 $g(\zeta)$:

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

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

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 [60], 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 and 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. On each HOG feature we apply a max-filtration using a 5x5 kernel. The max-filter emphasizes and expends the HOG-filter responses to the neighboring pixels. Additionally, we add the output of a skin detector [61] and 6 color features, which are obtained by applying max- and min-filter with a 5x5 kernel on each Lab color channel.

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 (see Figure 2). All negative patches have at least 0.2 of the average upper body size distance to the body part annotation. Each patch P is therefore augmented by a 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 [60]:

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

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 [42], [58]. The influence of the averaging is well discussed in [62].

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

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

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. [60]. The impact of different optimization schemes with two objectives has been evaluated in [52].

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 Figure 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:

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

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 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

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

5 Implementation Details

There are different ways to implement the binary potentials $\psi_{kl}(\mathbf{j}_k, \mathbf{j}_l)$ in the pictorial structure model discussed in Section 3. As already mentioned, our approach uses a joint representation and from the training data we obtain the relative joint positions of a child with respect to its parent in the tree model. Two examples are shown in Figure 3. The simplest approach models the relative positions by a Gaussian distribution with mean μ_{kl} and covariance matrix Σ_{kl}. As in [63], the binary potential becomes

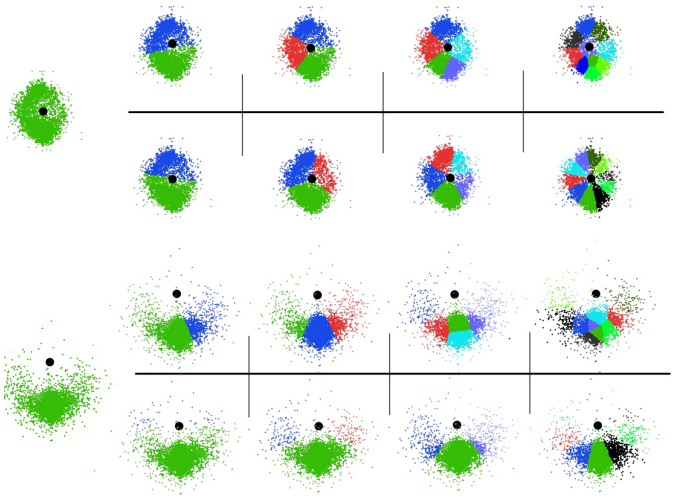


Fig. 3. We model the spatial relation of the joints by clustering the relative offset vectors of the joints with respect to their parents (*black dot*). The top two rows show the distribution of the locations of the left hand with respect to the left elbow. The two rows at the bottom display the distribution of the locations of the left knee with respect to the left hip. This distribution has a bigger bias towards one direction. For each joint, the two rows show the difference between clustering of the offset vectors (*first row*) and uniform quantization of the limb orientations (*second row*) for a varying number of clusters or bins. The impact of the clustering on the performance is plotted in Figure 6.

$$\psi_{kl}(\mathbf{j}_k, \mathbf{j}_l) = \exp(-\frac{1}{2}((\mathbf{j}_l - \mathbf{j}_k) - \mu_{kl})^T \Sigma_{kl}^{-1}((\mathbf{j}_l - \mathbf{j}_k) - \mu_{kl})). \quad (14)$$

Σ_{kl} can be further assumed to be a diagonal matrix, which can be achieved by a singular value decomposition [63]. While a Gaussian distribution is a good model for a limb representation that explicitly models the orientation and length of each limb as in [63], Figure 3 shows that a Gaussian distribution is a rather poor model for a joint representation. We therefore follow the data-driven approach proposed in [64] and cluster the relative locations using a standard k-means algorithm. Each cluster m for a joint pair $(\mathbf{j}_k, \mathbf{j}_l)$ is then modeled by a Gaussian with μ_{kl}^m and Σ_{kl}^m. While Figure 3 visualizes the difference between clustering the relative positions and a uniform quantization of the limb orientations, a quantitative evaluation is given in Section 7.

Inference is performed as in [64], where we start from the leaves of the tree model and move towards the root node. Since a joint can be assigned only to one cluster m, we take only the best cluster before moving to the parent. We also take the cluster frequency within the training

data into account by weighting the clusters:

$$w_{kl}^m \exp(-\frac{1}{2}((\mathbf{j}_l - \mathbf{j}_k) - \mu_{kl}^m)^T (\Sigma_{kl}^m)^{-1}((\mathbf{j}_l - \mathbf{j}_k) - \mu_{kl}^m)), \quad (15)$$

where $w_{kl}^m = p(m|k,l)^\alpha$. In the experiments, we found that the setting $\alpha = 1$ penalized clusters that occur rarely too much and therefore downweighted the cluster probability by $\alpha = 0.1$.

While in [16] a single pictorial structure model was used, we also extend the approach by having several models as in [17]. To this end, we represent the poses by a 24×2 dimensional feature vector containing for each joint except of the head the relative 2d offset vector to its parent and also to the head. The clustering is then performed by a k-means algorithm. Examples of the pose space clustering are shown in Figure 8. For pose estimation, we do the inference for each pictorial structure model and take the one with the highest score.

In contrast to [17], we do not learn different part templates for each model but use the same parts dependent joint regressors for all models. This has the advantage that the additional time required for training can be neglected and the testing time only slightly increases with the number of models as we show in Section 7. Although it has been shown in [17] that it is beneficial to weight the different models, we use uniform weights in our experiments.

6 FashionPose Dataset

Fig. 4. 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, corresponding to 10%, 15%, and 25% of the upper body size.

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 [65], [17] since these two points were very difficult to annotate accurately. Occluded joints have also been annotated. Some example images with ground truth annotation are shown in Figure 4.

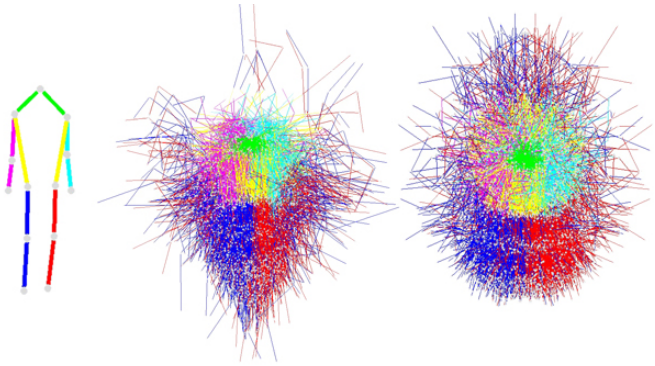


Fig. 5. The scatter plots visualize the pose variability over the datasets FashionPose (center) and Leeds Sports Pose (LSP) [17] (right). The single skeleton on the left hand side explains the color code of the different body parts. The skeletons are centered on the nose, which is the top joint, and the skeletons of FashionPose are rescaled based on the upper body size. While the LSP dataset has a higher variation in poses and includes also a significant number of persons in an upside-down pose, the FashionPose dataset has a higher variation in limb length even in the scale normalized version.

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. For evaluation, we use two versions of the dataset, namely a scale normalized version as in [16] and an unscaled version. For scale normalization, we use the distance between the average position of the two hip joints and the average position of the two shoulder joints and rescaled all images to a common upper body size of 75 pixels. Table 1 shows a detailed analysis of the distances between the body joints.

The dataset is more challenging than the Fashionista dataset [65] 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. In comparison to the well-known Leeds Sports Pose dataset (LSP) [17], the dataset contains less

pose variation as shown in Figure 5, but a much higher variation of appearance and dress style, which is rather small within each of the eight sport classes in [17]. The FashionPose dataset is publicly available.[1]

	rescaled images mean (std)	original images mean (std)
head - shoulder	32.00 (8.83)	102.93 (41.53)
shoulder - elbow	41.97 (7.78)	134.67 (48.02)
elbow - hand	33.37 (9.23)	106.20 (42.09)
shoulder - hip	75.98 (3.34)	246.95 (79.73)
hip - knee	54.65 (12.62)	163.69 (50.74)
knee - feet	54.35 (12.48)	158.23 (49.65)

TABLE 1
All 7,543 annotations have been rescaled to common upper body size of 75 pixel. The mean and the standard deviation of the distances in pixels between body joints with and without normalization are shown.

7 Experiments

For a quantitative and a qualitative evaluation we use two datasets, namely the well-known Leeds Sports Pose dataset (LSP) [17] and the newly collected *FashionPose* dataset described in Section 6. All images in the LSP dataset are rescaled such that the most prominent person in the image is roughly 150 pixel in scale, yielding an average upper body size of 43.31 pixel. In our experiments, we compare our method to several methods, namely linear and non-linear SVMs for part templates [17], [66], flexible mixtures-of-parts [5], spatial hierarchies of mixture models [36], pictorial structures with appearance sharing [21], and the recently published poselet conditioned pictorial structures [38], [39], visual symbols [41], and latent tree models [29].

The evaluation on the FashionPose dataset has been performed on two different versions, once we took advantage of the ground truth and rescaled all the images to an upper body size of 75 pixel and once we used the images in original scale. On this dataset, we compare our approach to the flexible mixtures-of-parts approach [5].

7.1 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 [67]. PCP declares a limb as correctly detected if the error of both endpoints is within 50% of the limb length from the ground truth endpoints. Note that some works use a simplified version of PCP that results in much higher accuracy numbers. For instance, one measure takes the mean of the endpoints instead of the endpoints themselves as accuracy measure. Another measure uses the ground-truth upper body bounding box for evaluation. In order to be consistent, we use the strict PCP measurement for comparison with other reported results on the Leeds Sports Pose dataset; otherwise we use the more informative normalized joint localization error.

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

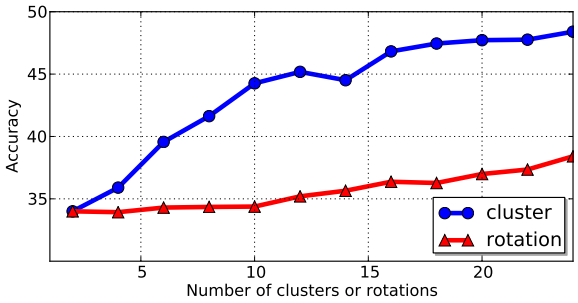
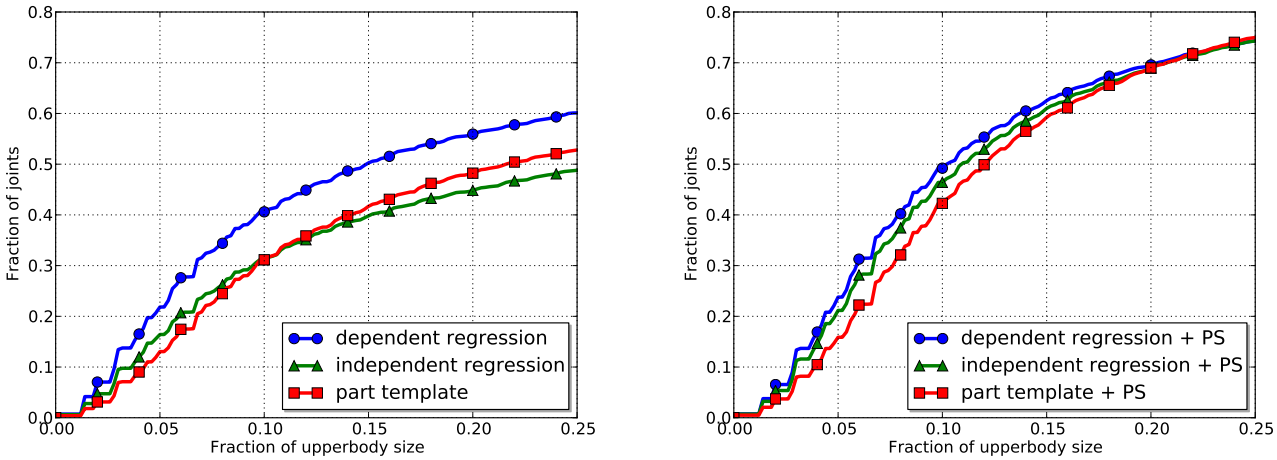


Fig. 6. Comparison of a data-driven approach, which clusters the relative location of the joints with respect to their parents, and a uniform quantization of the limb orientations. The performance increases for both approaches with the number of clusters or bins, but the clustering approach performs better since it also implicitly models foreshortening. Figure 3 visualizes the two approaches.

7.2 Experiments on FashionPose

Unless otherwise stated, we use the scale-normalized version of the dataset. For the training of the body part templates, independent and parts dependent joint regression, we use similar parameters as in [16]. The forests of all three different approaches consists of 15 trees with a maximum depth of 20 and a minimum number of 20 patches per leaf. For training, we generate $40,000$ binary tests (7) at each node, where we use $1,000$ random parameter settings for $\gamma\backslash\tau$ and for each setting additionally 40 random thresholds τ. Each tree has been grown on a set of $400,000$ positive and $400,000$ negative patches extracted from $4,000$ randomly selected training images. Please note that in [16] we used $500,000$ positive and negative patches, this change has a negligible influence on the final performance. The size of the extracted patches, and thus of the feature matrices F_P^f, is 60% of the upper-body size. For computational reasons, we evaluate the split functions at each node for only maximal 200,000 patches. Each forest and each tree can be trained independently, so the training of an entire layer of our approach takes approximately 3 hours on a 700 CPU cluster. In order to avoid overfitting of the two layer system, we mirrored the training images for the second layer.

Joint Regressors. 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 Figure 7 (a).



(a) (b)

Fig. 7. Accuracy plots for the three proposed methods without (a) and with (b) the pictorial structure model. The parts dependent joint regressors outperform the independent regressors and the part templates. Adding a pictorial structure model improves the performance for all three methods. While the differences become smaller, the performance of the parts dependent joint regressors is still higher then the other methods at lower error thresholds.

	avg	Head	Shoulder	Hip	Elbow	Wrist	Knee	Ankle
Without Pictorial Structure								
body part template	31.15	61.42	54.42 48.39	25.68 23.10	13.68 11.61	21.94 20.26	32.26 29.16	34.84 37.16
independent joint regression	31.20	70.97	45.81 47.10	23.74 20.26	10.71 09.68	17.81 17.94	37.55 34.06	36.00 33.94
dependent joint regression	40.61	76.65	57.55 57.55	30.06 30.71	22.58 20.00	28.77 24.77	45.68 43.74	43.35 46.45
With Pictorial Structure								
body part template	42.26	51.23	52.26 52.13	52.13 52.39	38.84 39.87	27.23 27.23	44.26 41.29	37.16 33.42
independent joint regression	46.10	64.52	61.55 58.45	57.29 53.16	41.68 38.84	28.65 27.35	46.06 42.84	41.16 37.81
dependent joint regression	49.21	69.16	62.97 61.81	60.90 58.84	38.32 40.13	31.35 28.26	50.45 49.42	43.23 44.90
With Mixture of Pictorial Structures								
dep. joint regression with 3 models	50.52	70.32	65.68 62.58	61.29 60.00	43.23 41.29	30.84 29.55	51.48 51.48	44.65 44.39
dep. joint regression with 7 models	51.28	69.16	64.77 64.90	61.29 60.65	45.16 40.13	31.87 28.77	54.19 55.10	45.94 44.65
dep. joint regression with 10 models	52.02	71.35	66.84 64.13	65.68 59.61	43.87 40.90	33.55 28.00	55.48 53.55	47.23 46.06

TABLE 2

Detailed detection accuracy for all joints at error threshold 0.1 (10% of the upperbody size). 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). The pictorial structure model improves the performance for all three cases. Having several pictorial structure models improves the performance further as shown in the last three rows.

The proposed body parts dependent joint regressors clearly outperform the independent part templates and joint regressors. In particular at low error rates, the joint regressors are more accurate then the part templates. The accuracy for each joint is provided in Table 2 (Without Pictorial Structure). In particular the accuracy of the ambiguous joints of the arms and legs is strongly improved.

Pictorial Structure. When using the part templates, the independent joint regressors, and the body parts dependent joint regressors within a pictorial structure framework (Section 3), the performance is increased due to the use of the prior knowledge about the human body. The accuracy of all three methods is given in Figure 7 (b) and Table 2 (With Pictorial Structure).

As discussed in Section 5, we investigated two ways of modeling the spatial relation of the joints within the pictorial structure model. The first approach is data driven and clusters the offset vectors of the joints with respect to their parents. The second approach uniformly quantizes the limb orientations. The difference between the two approaches is shown in Figure 3. The performance with respect to the number of clusters or bins is plotted in Figure 6. Since the clustering implicitly models variations in scale of the persons and the foreshortening of the limbs, clustering performs better than a uniform model for orientation. The plot also shows that the performance increases with the number of clusters. In the rest of the paper we use 20 clusters per each joint.

We also investigated the benefit of having several pictorial models as discussed in Section 5. Figure 8 shows some examples of the pose clustering. When learning a model for each pose cluster, the performance is improved as shown in Figure 9.

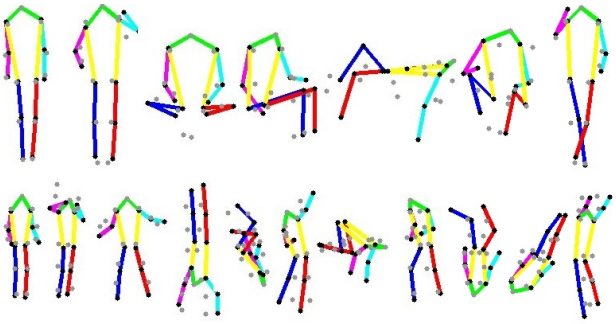


Fig. 8. The first row shows different clusters obtained by clustering the pose space for the FashionPose training set. The gray dots visualize the cluster centroids and the skeleton shows the nearest sample of the training set for each centroid. The skeletons represent also typical poses of the dataset. Most of the clusters represent different semantic meaningful poses, e.g., standing, sitting, laying. Some clusters, however, contain similar poses at different scales. The lower row shows some examples for the LSP training data. As in Figure 5, the clusters show the larger pose variation in the LSP dataset.

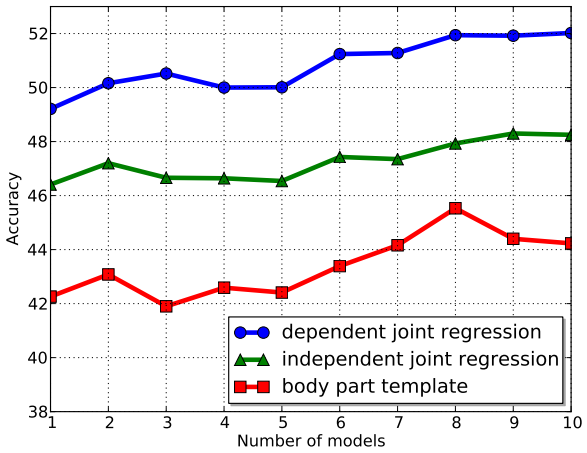


Fig. 9. Using multiple pictorial structure improves the performance for all three proposed methods.

Size of Patches. The size of the extracted patches has a big influence on the final performance a shown in Figure 10. For the part templates and the independent joint regressors, a larger patch size leads to better results, but not within a pictorial structure framework. The parts dependent joint regressors show a more stable performance in both cases and perform well for a patch size of 60-70% of the upper body size.

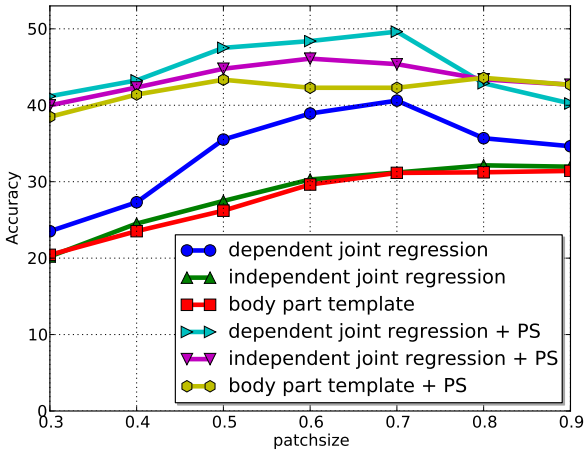


Fig. 10. The influence of the size of the extracted patches with respect to the size of the upper body. For the body part templates and the independent joint regressors, larger patches achieve better accuracy. Integrated in a pictorial structure framework, the performances however drop if the patches are getting too big. In contrast, the parts dependent joint regressors perform well for a patch size of 70% of the upper body size with and without a PS model.

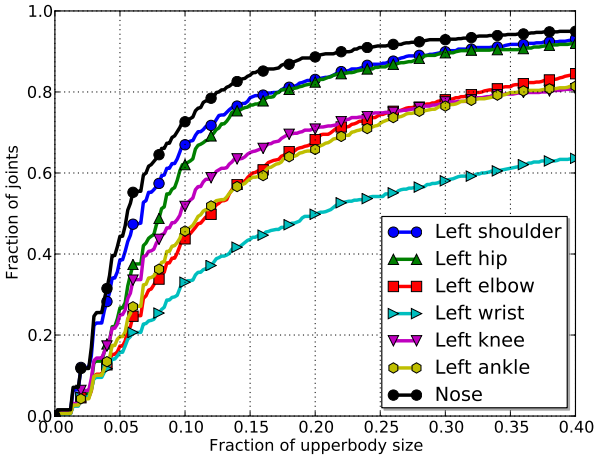


Fig. 11. The accuracy plots for individual joints using body parts dependent joint regressors with a pictorial structure model. For better readability, we plot only the left joints. As expected, localizing the wrist is the most difficult task, whereas head, shoulders, and hip joints are reasonable well localized. The numbers for all joints at error thresholds 0.1 and 0.15 (10% and 15% of the upper body size) are provided in Table 3.

Figure 12 gives an overview of the results achieved using the parts dependent joint regressors and a comparison with a state-of-the-art method proposed by Yang et al. [5] 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

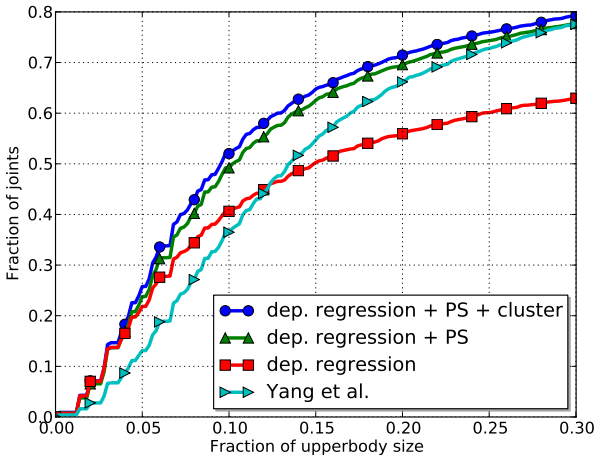


Fig. 12. Comparison of the joint localization accuracy of the proposed unary potentials and comparison with a state-of-the-art method [5]. Since the body parts dependent joint regressors do not encode any explicit information of the human skeleton, using a pictorial structure model (PS), which models the kinematic chain, gives an additional performance boost. Clustering the pose space and having several PS models gives an extra performance boost. The body parts dependent joint regression together with a pictorial structure model and the clustering outperforms [5] significantly. In particular at low error rates like 0.1, the number of correctly localized joints is 20% higher than [5].

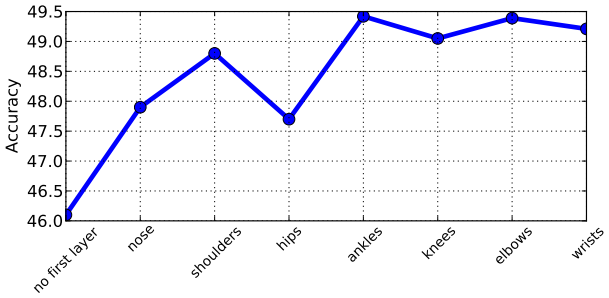


Fig. 13. Not all independent part potentials of the first layer have the same impact on the final results of the parts dependent joint regressors. The far left data point in this plot corresponds to independent joint regression and the point on the far right to the joint regressors that depend on all part potentials of the first layer. From left to right more parts are added where the order of parts is based on the accuracy of the potentials.

rescaled training images. The pictorial structure model with parts dependent joint regression outperforms [5]. There is a significant increase of estimates with a small error. Larger error thresholds indicate a poor accuracy that is probably insufficient for applications; see Figure 4. For error thresholds like 0.1, the accuracy is improved by more than 20%. Table 3 compares the accuracies for all joints at error thresholds 0.1 and 0.15. Our approach localizes the joints with a higher accuracy. It also compares the performance to [16] that uses only one pictorial structure model and slightly different training settings. The accuracy curves for individual joints are plotted in Figure 11.

We also evaluated the benefit of using more than two layers. For this experiment, we used the output of the previous two layers as input for the third layer. As first layer we used a classification forest and regression forests for the second and third layer. The use of the additional layer resulted in an accuracy increment by only $+0.3\%$. 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.

In order to evaluate the impact of the individual unary potentials of the first layer, we started with only image features as input for the regressor of the second layer, which is equal to the independent joint regression approach, and added the unary potentials of the first layer one by one. The adding order was defined by the final accuracy of the body joints. We started with the most accurate, namely the nose, and ended with the wrists. Figure 13 shows the performance change after adding more and more unary potentials of the first layer to the second layer. While adding only the most reliable part already improves the performance by 2%, the hips actually reduced the accuracy. Overall, it seems that not all parts are necessary for the first layer and a subset of parts like nose, shoulders, and ankles are sufficient.

Scale. So far we assumed that the location and scale of the person is known as in [16]. To test the abilities to handle multiple scales, we used the original images of the dataset and cropped the images if they contained multiple persons in the foreground. During test time all images are rescaled such that the longest side is maximal 512 pixel long. After the rescaling of the image, an image scale pyramid is built. The pyramid has 6 layers and the scale-factor from one layer to the next is 0.8. We determine the final pose estimation by simply taken the scale with the highest score. In our experiments we could see that the average detection accuracy at error threshold 0.1 drops only 3.16% to 48.86 when using the dependent joint regressors. Building the scale pyramid and extracting the features for each scale increased the computation by 10-15%. The computational overhead, however, can be reduced by using multiple threads.

7.3 Experiments on Leeds Sports Pose Dataset

The variation of poses is higher in the LSP dataset then in the novel FashionPose dataset as visualized in Figure 5,

error thres. joints	0.10 ours	0.10 [16]	0.10 Yang et al. [5]	0.15 ours	0.15 [16]	0.15 Yang et al. [5]
avg	52.02	49.16	37.34	64.75	62.89	55.26
Head	71.35	66.97	56.16	81.68	78.84	77.76
L. shoulder	66.84	61.94	53.21	78.06	73.81	72.75
R. shoulder	64.13	61.81	55.39	75.23	74.19	74.03
Left hip	65.68	57.16	38.43	76.52	72.90	58.61
Right hip	59.61	58.58	34.96	75.48	73.81	58.09
Left elbow	43.87	41.81	27.89	60.26	56.00	46.14
Right elbow	40.90	41.29	32.51	55.74	58.84	50.64
Left wrist	33.55	32.26	24.29	43.48	44.26	38.17
Right wrist	28.00	29.68	21.72	37.68	39.48	33.16
Left knee	55.48	52.13	39.07	66.45	65.29	56.94
Right knee	53.55	49.94	38.43	67.23	62.71	57.32
Left ankle	47.23	43.87	32.26	61.03	58.97	49.61
Right ankle	46.06	41.68	31.10	62.97	58.58	48.20

TABLE 3
Detection accuracy for all joints at error thresholds 0.1 and 0.15. The comparison shows that our method performs similar or better than [5] for all joints. Clustering the pose space and having a mixture of pictorial structure models also improved the previous version [16].

but at the same time the variation in image quality and clothing style is much smaller. For the evaluation of the LSP dataset [17], we trained 25 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. While the number of patches sampled per image is the same as before, the overall number of patches per tree is smaller since there are less training images. Since the test images are also smaller, we increased the number of trees to keep the test time roughly the same. In order to compare with previous works, we stick to the original 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. We are using the observer-centric annotation [21] approach, i.e., we flip the labels 'left' and 'right' such that the left limbs are always on the left side of the torso according to the shoulder and hips annotations.

A detailed comparison between our three methods is given in Table 4. Also on this dataset, we can see that the parts dependent joint regression algorithm outperforms the other two approaches regardless of using a pictorial structure. The parts dependent joint regression achieves a 10% higher score on the PCP criteria without taking advantage of the pictorial structure, this is mainly due to the more accurate estimation of the more challenging joints like the elbows and the knees. Besides of the original PCP criteria, we also evaluated the accuracy with respect to the normalized joint localization error for individual joints. The results are plotted in Figure 14. It shows that the performance varies a lot between the different joints.

The random forests provide two convenient parameters for finding an optimal trade-off between runtime and accuracy, namely the number of trees and the sampling stride, i.e., the distance between patches sampled at test time. In our experiments, we sample the patches very densely by using a sampling stride of two pixels. Such a high sampling rate is crucial for the body part templates, but in our experiments we saw that the regressors could handle lower sampling rates. As shown in Figure 15 (a), a higher number of trees improves the accuracy at the cost of a higher average computation time. We noticed a saturation at 30 trees indicating that the trees are correlated due to the limited size of the dataset. In fact, each tree has been trained on the same $1,000$ training images.

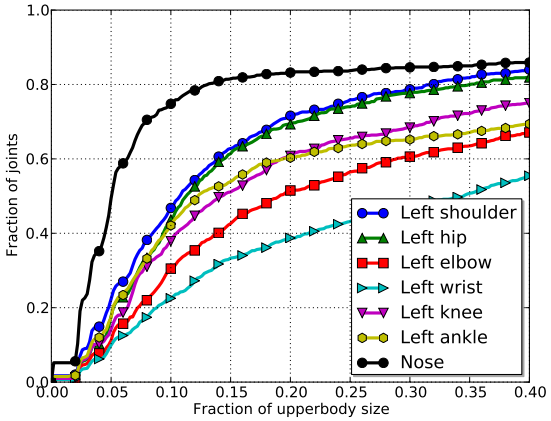


Fig. 14. Accuracy plots of the parts dependent joint regressors for individual joints on the LSP dataset.

Figure 15 (b) shows the influence of the pose space clustering and having several pictorial structure models. Example pose clusters are shown in Figure 8. A higher number of clusters leads to a better performance, but

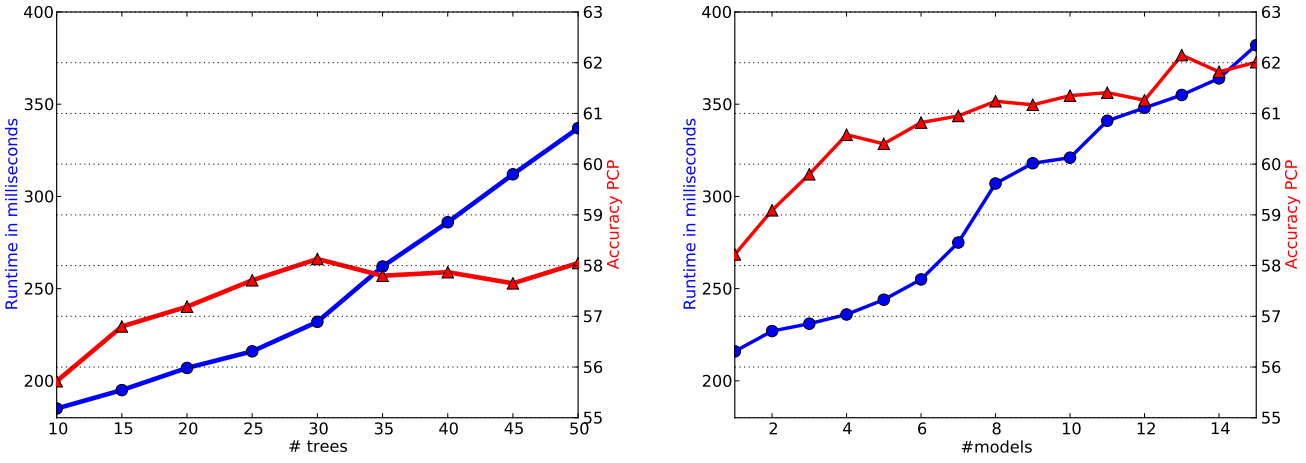


Fig. 15. Trade-off between runtime and accuracy. More trees (a) and more mixture models (b) increase the accuracy but at the same time also the runtime. The accuracy saturates for 30 trees. A similar saturation effect can be seen in plot (b); the first four mixture models give the best accuracy gain with respect to runtime. Additional models still improve the accuracy, but the increase becomes smaller. The plots also show that the additional runtime cost by additional trees and models are modest.

solving the inference of more pictorial structures also reduces the runtime. Since we only cluster the pose space in order to create multiple pictorial structures but do not train different regressors for each cluster, the additional training time can be neglected and the increase in runtime is moderate. The full system including the feature extraction, evaluation of 700 trees (25 trees per forest, 14 joints, and 2 layers), and inferences with one pictorial structure model takes 280 milliseconds on average[2].

Table 4 compares our approach with related methods and different state-of-the-art methods on this dataset. The comparison with a pictorial structure model that uses linear SVMs [17] or a cascade of non-linear SVMs [17] as part templates shows that all of our proposed unary potentials achieve a much higher accuracy.

In [17], the pose space has also been partitioned into 4 clusters to train an individual model for each cluster. As can be seen from Table 4, 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. In our approach, we only partition the pose space in order to learn multiple pictorial structures and we do not train individual appearance models for each cluster. Nevertheless, our approach already achieves better results by using only one appearance-model for each cluster. We also compare our method to the work [66], that uses 10,000 additional annotated training samples.

While our approach significantly outperforms Yang et al. [5] on the FashionPose dataset, the difference between these two methods is small on this dataset. The smaller difference can be partially explained by the PCP measure that does not evaluate joint localization accuracy, but limb detection performance. PCP tolerates a relatively high localization error, where the accuracy differences between our approach and [5] are also smaller on the FashionPose dataset. Figure 16 shows the PCP accuracy of both methods given different thresholds.

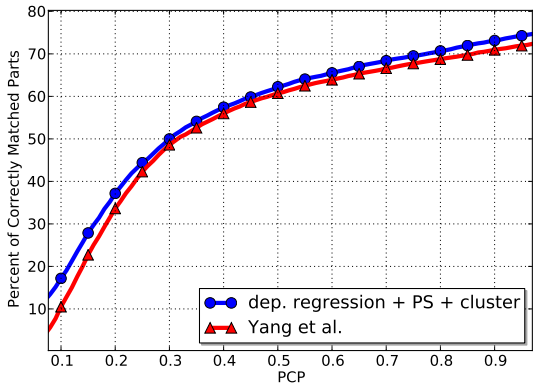


Fig. 16. PCP accuracy given different thresholds.

[36] uses a more complex model than a tree structure that captures the space of plausible human poses much better. While this method achieves comparable results on this dataset, they have probably to deal with higher cost for training and running times. [21] extends [5] by improving the hard negative mining and by using additional information about the color statistics of the dataset. The used color statistics are particularly helpful on the LSP dataset because it contains eight different sports classes where the foreground and background

2. Measured on 500 randomly selected images from the test set using Intel Core i7 3.06GHz with 4 cores (multi-threaded).

color is very similar within each sport class. Three recent methods achieve comparable [29], [38] or a slightly better [39], [41] accuracy compared to our approach.

Fig. 17. Some typical failure cases on the FashionPose and the LSP dataset.

Figures 17, 18, and 19 show some pose estimates including failure cases on the FashionPose and the LSP dataset.

8 Conclusion

In this work, we have addressed the problem of human pose estimation from still images within a pictorial structure framework. In our experiments, we have shown that random forests are an efficient and powerful tool for estimating unary potentials for this task. To overcome the limitations of independent part templates, we proposed dependent joint regressors that consist of random forests that operate over two layers. The first layer acts as a traditional independent body part classificator. The second layer does not only take the image features for estimating the joint locations into account but also the predicted distributions of the first layer, thus allowing to put the body parts into relation. In our experiments, we have shown that our proposed approach for estimating unaries outperforms independent body part classifiers and independent joint regressors within or without a pictorial structure framework.

We also proposed the novel dataset $FashionPose$, which complements existing datasets like the Leeds Sports Pose (LSP) dataset for pose estimation. The dataset exceeds existing datasets in terms of appearance variation due to the large variation of dressing styles. Our method achieves a PCP measure that is better or equally to state-of-the-art methods on the LSP dataset, however, it requires less than one second per image. Using a joint localization error instead of PCP on $FashionPose$ revealed that the proposed method performs very well for accurate joint localization.

So far we are assuming that only one person is present in the image. Detecting multiple persons in one image has a negligible impact on the computational time and can be implemented as in [68].

Acknowledgments

The authors acknowledge financial support from the ERC Grant (VarCity), the EC projects RADHAR (FP7-ICT-248873) and TANGO (FP7-ICT-249858), the CTI project (12618.1 PFES-ES), and the DFG Emmy Noether program (GA 1927/1-1).

References

[1] J. Shotton, R. Girshick, A. Fitzgibbon, T. Sharp, M. Cook, M. Finocchio, R. Moore, P. Kohli, A. Criminisi, A. Kipman, and A. Blake, "Efficient human pose estimation from single depth images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 12, pp. 2821–2840, 2013.

[2] Y. Liu, J. Gall, C. Stoll, Q. Dai, H.-P. Seidel, and C. Theobalt, "Markerless motion capture of multiple characters using multiview image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 11, pp. 2720–2735, 2013.

[3] M. A. Fischler and R. A. Elschlager, "The representation and matching of pictorial structures," *IEEE Transactions on Computers*, vol. 22, no. 1, pp. 67–92, 1973.

[4] P. F. Felzenszwalb and D. P. Huttenlocher, "Pictorial structures for object recognition," *International Journal of Computer Vision*, vol. 61, no. 1, pp. 55–79, 2005.

[5] Y. Yang and D. Ramanan, "Articulated human detection with flexible mixtures-of-parts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 12, pp. 2878–2890, 2013.

[6] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2005, pp. 886–893.

[7] X. Zhu, C. Vondrick, D. Ramanan, and C. C. Fowlkes, "Do we need more training data or better models for object detection?" in *British Machine Vision Conference*, 2012.

[8] 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 *IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 1297–1304.

[9] R. Girshick, J. Shotton, P. Kohli, A. Criminisi, and A. Fitzgibbon, "Efficient regression of general-activity human poses from depth images," *International Conference on Computer Vision*, pp. 415–422, 2011.

[10] L. Sigal and M. Black, "Measure locally, reason globally: Occlusion-sensitive articulated pose estimation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2006, pp. 2041–2048.

[11] X. Ren, A. C. Berg, and J. Malik, "Recovering human body configurations using pairwise constraints between parts," in *International Conference on Computer Vision*, 2005, pp. 824–831.

[12] H. Jiang and D. Martin, "Global pose estimation using non-tree models," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[13] T.-P. Tian and S. Sclaroff, "Fast globally optimal 2d human detection with loopy graph models," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 81–88.

[14] M. Bergtholdt, J. H. Kappes, S. Schmidt, and C. Schnörr, "A study of parts-based object class detection using complete graphs," *International Journal of Computer Vision*, vol. 87, no. 1-2, pp. 93–117, 2010.

Limps	Avg.	Head	Torso	Upper Leg		Lower Leg		Upper Arm		Forearm	
Without Pictorial Structure											
body part template	37.3	80.6	53.7	44.3	41.9	42.6	43.5	25.3	22.3	10.7	09.0
independent joint regression	33.4	80.3	50.6	38.5	37.5	35.4	35.2	19.7	18.1	10.1	08.9
dependent joint regression	44.8	82.3	66.5	50.9	50.3	49.2	51.3	33.4	34.9	14.6	14.6
With Pictorial Structure											
Linear SVM [17]	36.4	49.9	64.1	42.4	43.1	41.2	40.7	26.2	23.7	16.5	15.7
Non-linear SVM [17]	44.7	55.9	70.9	53.5	58.7	49.3	47.4	37.1	29.1	26.8	18.8
body part template	56.1	80.7	81.1	66.6	65.6	60.6	62.5	47.2	45.3	28.0	23.6
independent joint regression	55.1	81.5	81.0	63.5	63.6	57.6	58.2	47.2	46.6	27.2	24.6
dependent joint regression	58.4	83.3	83.4	68.4	69.7	63.7	62.9	48.3	47.7	27.9	25.1
With Mixture of Pictorial Structures based on Pose Clustering											
Linear SVM [17] + Clustering	43.6	59.7	74.1	54.4	53.6	49.0	49.3	30.5	30.9	17.5	17.7
Non-linear SVM [17] + Clustering	55.1	62.9	78.1	64.8	66.7	60.3	57.3	48.3	46.5	34.5	31.2
SVM (10000 images) [66] + Clustering	62.7	59.7	88.1	75.2	73.8	66.7	66.3	53.0	54.4	36.1	38.9
dependent joint regression + Clustering	62.2	86.2	86.0	73.3	72.6	66.7	66.2	54.8	52.8	33.6	29.3
State-of-the-art											
Yang et al. [5]	60.8	77.1	84.1	69.5		56.6		52.5		35.9	
Spatial Hierarchy of Mixture Models [36]	58.8	86.5	93.7	68.0		57.8		49.0		29.2	
Cluster + S. Hierarchy of M. M. [36]	61.3	57.8	95.8	69.9		60.0		51.9		32.9	
Wang et al. [29]	62.8	86.0	91.9	74.0		69.8		48.9		32.2	
Wang et al. [41]	65.2	84.7	92.2	78.1		67.5		54.7		37.2	
Eichner et al. [21]	64.3	80.1	86.2	74.3		69.5		56.5		37.4	
Pishchulin et al. [38]	62.9	78.1	87.5	75.7		68.0		54.2		33.9	
Pishchulin et al. [39]	69.2	85.6	88.7	78.8		73.4		61.5		44.9	

TABLE 4
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 using a single tree model or mixture of models.

[15] D. Tran and D. Forsyth, "Improved human parsing with a full relational model," in *European Conference on Computer Vision*, 2010, pp. 227–240.

[16] M. Dantone, J. Gall, C. Leistner, and L. Van Gool, "Human pose estimation using body parts dependent joint regressors," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3041–3048.

[17] S. Johnson and M. Everingham, "Clustered pose and nonlinear appearance models for human pose estimation." in *British Machine Vision Conference*, 2010.

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

[19] R. Ronfard, C. Schmid, and B. Triggs, "Learning to parse pictures of people," in *European Conference on Computer Vision*, 2002, pp. 700–714.

[20] V. K. Singh, R. Nevatia, and C. Huang, "Efficient inference with multiple heterogeneous part detectors for human pose estimation," in *European Conference on Computer Vision*, 2010, pp. 314–327.

[21] M. Eichner and V. Ferrari, "Appearance sharing for collective human pose estimation," in *Asian Computer Vision Conference*, 2012, pp. 138–151.

[22] S. Zuffi, O. Freifeld, and M. J. Black, "From pictorial structures to deformable structures," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 3546–3553.

[23] L. Pishchulin, A. Jain, M. Andriluka, T. Thormählen, and B. Schiele, "Articulated people detection and pose estimation: Reshaping the future," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 3178–3185.

[24] M. Burenius, J. Sullivan, and S. Carlsson, "3d pictorial structures for multiple view articulated pose estimation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3618–3625.

[25] M. Andriluka, S. Roth, and B. Schiele, "Discriminative appearance models for pictorial structures," *International Journal of Computer Vision*, vol. 99, no. 3, pp. 259–280, 2012.

[26] M. Sun, M. Telaprolu, H. Lee, and S. Savarese, "An efficient branch-and-bound algorithm for optimal human pose estimation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 1616–1623.

[27] A. Lehrmann, P. Gehler, and S. Nowozin, "A non-parametric bayesian network prior of human pose," in *IEEE International Conference on Computer Vision*, 2013.

[28] Y. Wang and G. Mori, "Multiple tree models for occlusion and spatial constraints in human pose estimation," in *European Conference on Computer Vision*, 2008, pp. 710–724.

[29] F. Wang and Y. Li, "Beyond physical connections: Tree models in human pose estimation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 596–603.

[30] B. Sapp, C. Jordan, and B. Taskar, "Adaptive pose priors for pictorial structures," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 422–429.

[31] L. Bo and C. Sminchisescu, "Twin gaussian processes for structured prediction," *International Journal of Computer Vision*, vol. 87, no. 1-2, pp. 28–52, 2010.

[32] D. Ramanan, "Learning to parse images of articulated bodies," in *Neural Information Processing Systems Conference*, 2006, pp. 1129–1136.

[33] M. Eichner, M. Marin-Jimenez, A. Zisserman, and V. Ferrari, "2d articulated human pose estimation and retrieval in (almost) unconstrained still images," *International Journal of Computer Vision*, vol. 99, no. 2, pp. 190–214, 2012.

[34] M. Sun and S. Savarese, "Articulated part-based model for joint object detection and pose estimation," in *International Conference on Computer Vision*, 2011, pp. 723–730.

[35] Y. Wang, D. Tran, and Z. Liao, "Learning hierarchical poselets for human parsing." in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 1705–1712.

[36] Y. Tian, C. L. Zitnick, and S. G. Narasimhan, "Exploring the spatial hierarchy of mixture models for human pose estimation," in *European Conference on Computer Vision*, 2012, pp. 256–269.

[37] L. Bourdev, S. Maji, T. Brox, and J. Malik, "Detecting people using

Fig. 18. Qualitative results on some representative images of the FashionPose dataset.

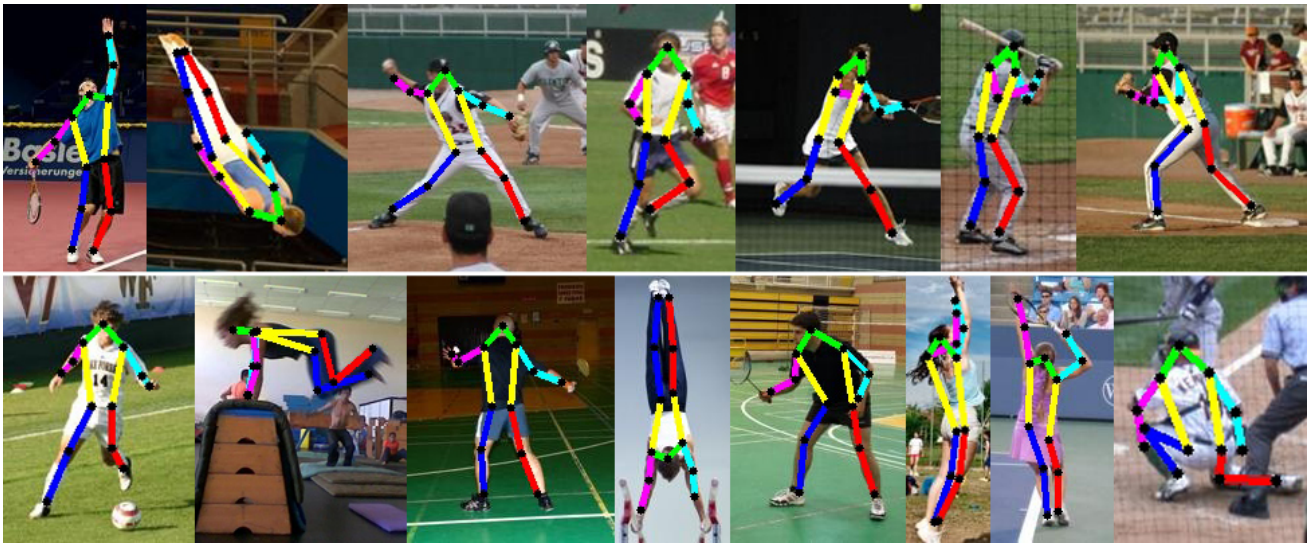


Fig. 19. Qualitative results on some representative images of the LSP dataset.

mutually consistent poselet activations," in *European Conference on Computer Vision*, 2010, pp. 168–181.

[38] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele, "Poselet conditioned pictorial structures," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 588–595.

[39] ——, "Strong appearance and expressive spatial models for human pose estimation," in *IEEE Intenational Conference on Computer Vision*, 2013.

[40] G. Gkioxari, P. Arbeláez, L. Bourdev, and J. Malik, "Articulated pose estimation using discriminative armlet classifiers," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3342–3349.

[41] F. Wang and Y. Li, "Learning visual symbols for parsing human poses in images," in *International Joint Conference on Artificial Intelligence*, 2013.

[42] Y. Amit, D. Geman, and K. Wilder, "Joint induction of shape features and tree classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 11, pp. 1300–1305, 1997.

[43] G. Rogez, J. Rihan, S. Ramalingam, C. Orrite, and P. H. Torr, "Randomized trees for human pose detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[44] G. Rogez, J. Rihan, C. Orrite, and P. H. Torr, "Fast human pose detection using randomized hierarchical cascades of rejectors," *International Journal of Computer Vision*, vol. 99, no. 1, pp. 25–52, 2012.

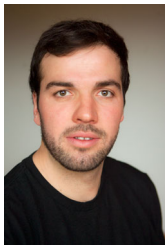
[45] B. Leibe, A. Leonardis, and B. Schiele, "Robust object detection with interleaved categorization and segmentation," *International Journal of Computer Vision*, vol. 77, no. 1-3, pp. 259–289, 2008.

[46] 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*, 2010, pp. 9–14.

[47] 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*, 2011, pp. 1196–1201.

[48] E. Sangineto, M. Cristani, A. Del Bue, and V. Murino, "Learning

discriminative spatial relations for detector dictionaries: an application to pedestrian detection," in *European Conference on Computer Vision*, 2012, pp. 273–286.
[49] D. Tang, Y. Liu, and T.-K. Kim, "Fast pedestrian detection by cascaded random forest with dominant orientation templates." in *British Machine Vision Conference*, 2012, pp. 1–11.
[50] C. Keskin, F. Kıraç, Y. E. Kara, and L. Akarun, "Hand pose estimation and hand shape classification using multi-layered randomized decision forests," in *European Conference on Computer Vision*, 2012, pp. 852–863.
[51] J. Shotton, M. Johnson, and R. Cipolla, "Semantic texton forests for image categorization and segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.
[52] G. Fanelli, M. Dantone, J. Gall, A. Fossati, and L. Van Gool, "Random forests for real time 3d face analysis," *International Journal of Computer Vision*, vol. 101, no. 3, pp. 437–458, 2013.
[53] M. Dantone, J. Gall, G. Fanelli, and L. Van Gool, "Real-time facial feature detection using conditional regression forests," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2578–2585.
[54] T. F. Cootes, M. C. Ionita, C. Lindner, and P. Sauer, "Robust and accurate shape model fitting using random forest regression voting," in *European Conference on Computer Vision*, 2012, pp. 278–291.
[55] B. Glocker, O. Pauly, E. Konukoglu, and A. Criminisi, "Joint classification-regression forests for spatially structured multi-object segmentation," in *European Conference on Computer Vision*, 2012, pp. 870–881.
[56] N. Razavi, J. Gall, P. Kohli, and L. Van Gool, "Latent hough transform for object detection," in *European Conference on Computer Vision*, 2012, pp. 312–325.
[57] D. Tang, T.-H. Yu, and T.-K. Kim, "Real-time articulated hand pose estimation using semi-supervised transductive regression forests," in *IEEE International Conference on Computer Vision*, 2013.
[58] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
[59] A. Criminisi and J. Shotton, *Decision Forests for Computer Vision and Medical Image Analysis*. Springer, 2013.
[60] J. Gall, A. Yao, N. Razavi, L. Van Gool, and V. Lempitsky, "Hough forests for object detection, tracking, and action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 11, pp. 2188–2202, 2011.
[61] M. Jones and J. Rehg, "Statistical color models with application to skin detection," *International Journal of Computer Vision*, vol. 46, no. 1, pp. 81–96, 2002.
[62] A. Criminisi and J. Shotton, "Decision forests for computer vision and medical image analysis," 2013.
[63] P. F. Felzenszwalb, R. B. Girshick, D. A. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.
[64] Y. Yang and D. Ramanan, "Articulated pose estimation with flexible mixtures-of-parts," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 1385–1392.
[65] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, and T. L. Berg, "Parsing clothing in fashion photographs," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 3570–3577.
[66] S. Johnson and M. Everingham, "Learning effective human pose estimation from inaccurate annotation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 1465–1472.
[67] V. Ferrari, M. Marin-Jimenez, and A. Zisserman, "Progressive search space reduction for human pose estimation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.
[68] M. Andriluka, S. Roth, and B. Schiele, "Pictorial structures revisited: People detection and articulated pose estimation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 1014–1021.



Matthias Dantone received the diploma degree in media informatics from the Ludwig Maximilian University of Munich in 2010. He is currently a PhD candidate at the Computer Vision Laboratory at ETH Zurich, where he works on topics related to human analysis, such as body pose estimation, face analysis and apparel recognition.

Juergen Gall obtained his B.Sc. and his Masters degree in mathematics from the University of Wales Swansea (2004) and from the University of Mannheim (2005). In 2009, he obtained a Ph.D. in computer science from the Saarland University and the Max Planck Institut für Informatik. He was a postdoctoral researcher at the Computer Vision Laboratory, ETH Zurich, from 2009 until 2012 and senior research scientist at the Max Planck Institute for Intelligent Systems in Tübingen from 2012 until 2013. Since 2013, he is professor at the University of Bonn and head of the Computer Vision Group.

Christian Leistner is a researcher and software engineer at Bing's Geospatial Division, Microsoft. He holds an MSc and PhD degree of Graz University of Technology. His recent research interests focus on large-scale machine learning and computer vision problems. He is particularly interested in object recognition, classification and semantic segmentation.

Luc Van Gool got a degree in electromechanical engineering at the Katholieke Universiteit Leuven in 1981. Currently, he is professor at the Katholieke Universiteit Leuven in Belgium and the ETH in Zurich, Switzerland. He leads computer vision research at both places, where he also teaches computer vision. He has authored over 200 papers in this field. He has been a program committee member of several major computer vision conferences. His main interests include 3D reconstruction and modeling, object recognition, tracking, and gesture analysis. He received several Best Paper awards. He is a co-founder of 5 spin-off companies.