Poselet Conditioned Pictorial Structures

Leonid Pishchulin[1] Mykhaylo Andriluka[1] Peter Gehler[2] Bernt Schiele[1]

[1]Max Planck Institute for Informatics, Saarbrücken, Germany

[2]Max Planck Institute for Intelligent Systems, Tübingen, Germany

Abstract

In this paper we consider the challenging problem of articulated human pose estimation in still images. We observe that despite high variability of the body articulations, human motions and activities often simultaneously constrain the positions of multiple body parts. Modelling such higher order part dependencies seemingly comes at a cost of more expensive inference, which resulted in their limited use in state-of-the-art methods. In this paper we propose a model that incorporates higher order part dependencies while remaining efficient. We achieve this by defining a conditional model in which all body parts are connected a-priori, but which becomes a tractable tree-structured pictorial structures model once the image observations are available. In order to derive a set of conditioning variables we rely on the poselet-based features that have been shown to be effective for people detection but have so far found limited application for articulated human pose estimation. We demonstrate the effectiveness of our approach on three publicly available pose estimation benchmarks improving or being on-par with state of the art in each case.

1. Introduction

In this paper we consider the challenging task of articulated human pose estimation in monocular images. State-of-the-art approaches in this area [2, 15, 26] are based on the pictorial structures model (PS) and are composed of unary terms modelling body part appearance and pairwise terms between *adjacent* body parts and/or joints capturing their preferred spatial arrangement. While this approach leads to tree-based models and thus efficient and exact inference, it fails to capture important dependencies between *non-adjacent* body parts. That modelling such dependencies is important for effective pose estimation can be seen e.g. in Fig. 1: activities of people like playing soccer, tennis or volleyball results in strong dependencies between many if not all body parts; this can not be modelled with the above approach.

This well known problem has so far been addressed in two ways. The first simply uses a mixture of tree models thus learning separate pairwise terms for different global body configurations *e.g.* [14, 15]. The second approach is to add more pairwise terms including non-adjacent body parts leading to a loopy part graph that requires approximate inference [2, 23, 21, 25]. A key challenge in designing models for pose estimation is thus to encode the higher-order part dependencies while still allowing efficient inference. In this paper we propose a novel model that incorporates higher order information between body parts by defining a conditional model in which all parts are a-priori connected, but which becomes a tractable PS model once the mid-level features are observed. This allows to effectively model dependencies between non-adjacent parts while still allowing for exact and efficient inference in a tree-based model.

Clearly, the choice of the particular mid-level image representation used for conditioning our model is crucial for good performance of the overall approach. On the one hand, this representation has to be robust with respect to variations in people appearance, pose and imaging conditions. On the other hand, it has to be highly informative for the underlying human pose. In order to satisfy these requirements we rely on the non-parametric poselet representation introduced in [4]. Note that for the task of people detection the best performing approaches are those which rely on a representation that jointly models appearance of multiple body parts [4, 10]. Yet these models have not been shown to lead to state-of-the-art performance in human pose estimation, likely because they rely on a pose representation that is not fine-grained enough to enable localisation of all body joints.

Related work. Most recent methods for human pose estimation are based on the pictorial structures (PS) model [12, 11] that represents the body configuration as a collection of rigid parts and a set of pairwise part connections. The connections between parts are typically assumed to form a tree structure in order to allow efficient inference at test time. Yet, several recent approaches considered non-tree models that allow to capture cues such as appearance similarity between limbs [22, 23, 20]. With a few excep-

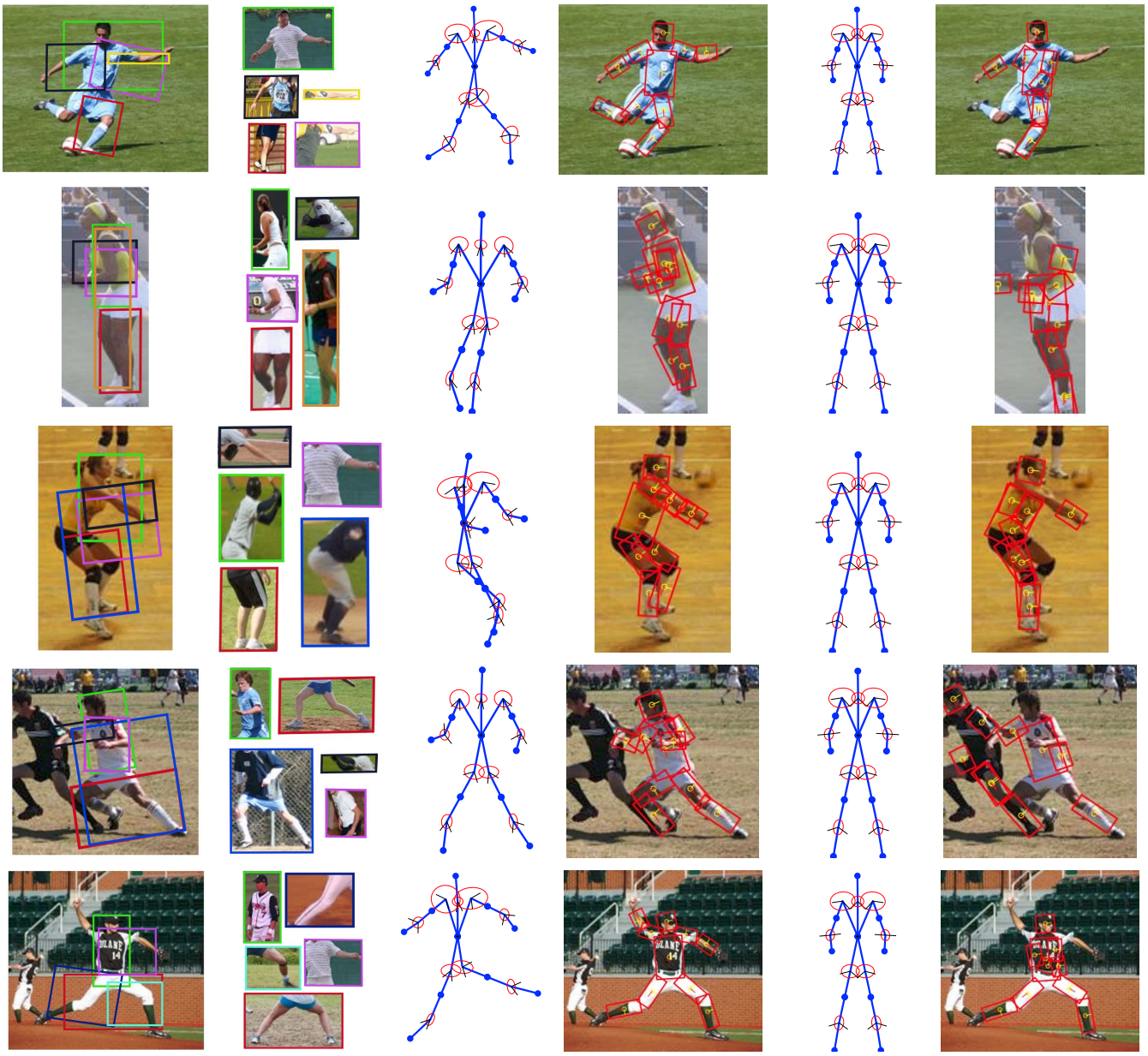

Figure 1. Visualisation of our approach. (a) shows the top scoring poselet detections with the corresponding poselet cluster medoids (b). It is visible that the poselets capture the anatomical configuration of the human in the input image. All poselet detections contribute to a prediction of the deformable pairwise terms, the outcome of which is shown in (c). Using the PS model with these pair-wise terms achieves the detection outcome (d). In contrast we show the generic prior [3] (e) and the corresponding pose prediction (f).

tions [21, 25] none of these models consider interactions between body parts that go beyond simple pairwise relationships. E.g. [25] proposes an approach that relies on a complex hierarchical model that requires approximate inference with loopy belief propagation. Our model is related to recent work aiming to increase the flexibility of the PS approach by jointly training a mixture of tree-structured PS models [14, 26, 7]. In particular, our model can be seen as an exponentially large collection of PS models with a selection function that chooses a suitable model based on the observed poselet features. Similar to these models, our approach allows efficient inference at test time, yet we are also able to incorporate dependencies between parts that go beyond pairwise interactions. Those are not captured in the model structure but in the conditioning step.

Our approach is related to holistic pose estimation ap-

proaches [1, 24, 18, 13] that aim to directly predict positions of body joints from image features without relying on an intermediate part-based representation. These methods demonstrate excellent performance in laboratory settings with little background clutter and are capable of recovering poses even in 3D. However, they have not been shown to perform well on real-world images with cluttered backgrounds considered in this work. Our approach is also related to [18] that aims to classify the image as a particular pose class. However, in our work we perform classification on the level of each body joint which allows the set of pose classes to be exponentially large.

In pose estimation literature the method of [19] is probably the closest to ours. Similar to their work, we define a PS model where unary and pairwise terms are image conditioned. However, our method is more general as it implicitly models dependencies between multiple parts by using an intermediate poselet-based feature representation. In contrast they rely on silhouette based similarity cues that are ineffective in the presence of background clutter, and act relatively local and thus capture mostly local pairwise part interactions. This makes our method applicable to more challenging sport images showing highly articulated humans from different viewpoints, while the method of [19] has been applied to frontal poses only with a comparatively small degree of articulation.

2. Review of Pictorial Structures

In this section we introduce the Pictorial Structures (PS) version [2, 3] that we are building on and that will serve as a baseline in the experiments. This implementation has been found to be competitive across a range of datasets. Although we focus on this particular incarnation of the PS model, we believe the extensions are applicable to other models, such as the one from [26]. The extension of this model will then be the topic of the next section.

We phrase the PS model as a conditional random field (CRF), modelling the conditional probability of a body pose configuration given image evidence. We denote by $L = (l_1, \ldots, l_M)$ a full body pose, consisting of M parts. A part $l_m = (x_m, y_m, \theta_m, s_m)^\top$ is parameterised by its x, y center position, rotation $\theta \in [0, 360)$, and scale $s \in \mathbb{R}_+$. With D we denote any form of image evidence and with β the vector of model parameters. For convenience we distinguish between parameters for unary β^u and pairwise β^p factors. The PS model then takes the form

$$E(L; D, \beta) = \sum_{m=1}^{M} E^u(l_m; D, \beta^u) + \sum_{n \sim m} E^p(l_n, l_m; \beta^p). \quad (1)$$

With $n \sim m$ we denote the neighbourhood relationship between the body parts. This typically is restricted to form a tree in order to enable exact and efficient inference.

Unary potentials We use the following unary potential functions

$$E^u(l_m; D, \beta^u) = \log \phi^u(l_m; D), \quad \forall m = 1, \ldots, M, \quad (2)$$

with pre-trained AdaBoost classifiers as the feature functions

$$\phi^u(l_m; D) = \max \left(\frac{\sum_t \alpha_i^t h_t(l_m, D)}{\sum_t \alpha_i^t}, \epsilon_0 \right). \quad (3)$$

A decision stump h_t in Eq.(3) is of the following form

$$h_t(l_m, D) = \text{sign}(\xi_t(\mathbf{x}_{n(t)} - \varphi_t)), \quad (4)$$

where $\mathbf{x}$ is a feature vector, $\varphi_t \in \mathbb{R}$ a threshold, $\xi_t \in \{-1, 1\}$, and $n(t)$ is a feature index. The feature vector is obtained by concatenating the shape context descriptors computed on a regular grid inside the part bounding box. We refer the reader to [2, 3] for details on training and descriptors.

Pairwise potentials Pairwise potential functions take the form

$$E^p(l_n, l_m; \beta^p) = \left\langle \beta^p_{n,m}, \phi^p_{n,m}(l_n, l_m) \right\rangle, \quad \forall n \sim m. \quad (5)$$

The features for the potential $\phi^p_{n,m}$ acting on n and m are computed as follows. First both parts are transformed into a common reference space, that is the location of the joint between these parts. We use the transformation

$$T_{mn}(l_n) = \begin{pmatrix} x_n + s_n \mu_x^{mn} \cos\theta_n - s_n \mu_y^{mn} \sin\theta_n \\ y_n + s_n \mu_x^{mn} \cos\theta_n - s_n \mu_y^{mn} \sin\theta_n \\ \theta_n + \tilde{\theta}_{mn} \\ s_n \end{pmatrix}, \quad (6)$$

where $\mu^{mn} = (\mu_x^{mn}, \mu_y^{mn})^T$ is the mean relative position of the joint between parts m and n in the coordinate system of part n; $\tilde{\theta}_{mn}$ is the relative angle between parts. The pairwise term is then a Gaussian on the difference vector between the two transformations $T_{mn}(l_n) - T_{nm}(l_m)$, as is standard practice in all PS works [2, 3, 26, 11]. We derive a linear form for the pairwise term in Eq. 5 using the natural parameterization of the Gaussian as in [10, 26], and place positivity constraints on those parameters in β^p that correspond to variances.

We learn unary and pairwise terms in a piecewise strategy, unary potentials using AdaBoost and the pairwise terms using a Maximum-Likelihood estimate.

3. Poselet Conditioned Pictorial Structures

Our approach is based on the following idea: we use a mid-level representation that captures possible anatomical configurations of a human pose to predict an image-specific

pictorial structures (PS) model that in turn is applied to the image. The representation we are using is inspired by the work [5, 25] which is why we refer to it as *poselets*. Poselets go beyond standard pairwise part-part configurations and capture the configuration of multiple body parts jointly. As we still predict a tree connected PS model we retain efficient and tractable inference.

The idea of our model is visualised in Fig. 1. On the input images we compute poselet responses that capture different portions of the person's body configuration. Highest scoring poselet detections are shown in Fig. 1(a), together with representative examples for them in Fig. 1(b). This information is then used to augment both unary and pairwise terms of the PS model. In Fig. 1(c) we show the deformation terms of the resulting PS model that we are able to predict. Pose of the person estimated with our poselet-conditioned model is shown in Fig. 1(d). For comparison we show the deformation model of [3] (a generic pose prior being the same for all images) along with the corresponding pose estimate in the last two columns.

The idea of having multiple deformation models is similar to the idea of encoding body pose configurations through different mixture components as in [26]. However, in their work the pairwise mixture components are – in contrast to our model – not dependent on the image but estimated during inference. We experimentally compare to this approach.

This section first describes the feature representation used to capture human poses. We then present the extension of the standard PS model outlined in the previous section and show how both unaries (sec. 3.2) and pairwise terms (sec. 3.3) can be enhanced using poselet information.

3.1. Poselet Representation

The goal of the mid-level representation is to capture common dependencies of multiple body parts. We implemented the following strategy to train a set of poselet detectors and compute a feature based on their responses.

For a reference body part, we cluster the relative positions of a subset of related body parts. For example, when picking the 'neck' part we cluster relative offsets of all upper body parts using Euclidean distance and K-means. We prune clusters that have less than 10 examples and use the remaining ones as poselets. In this paper we run this process multiple times, picking different reference points and multiple subsets of related parts to obtain a total of P clusters. Together with every poselet p we store its mean offset from the torso annotation μ_p.

The next step is to learn a detector for each poselet. Following [2, 3], we train AdaBoost detectors on dense shape context features. A separate detector is trained for every poselet cluster using all training images that fall within this cluster. Example outcomes can be seen in Figure 1(a+b) showing the highest scoring poselets for some sample images and their medoids.

To form a feature vector $f \in \mathbb{R}^P$ we first predict the torso position μ_{torso} in the test image. Given a torso prediction and the relative offset μ_p of the poselet p, we compute the maximum poselet response in a small region[1] around $\mu_{torso}+\mu_p$. This corresponds to a max-pooling step in a local region for every poselet p. Then we aggregate the maximum scores for all $p = 1, \ldots, P$ poselets to form a feature vector $f \in \mathbb{R}^P$. Similar to [25], we define 11 body part configurations, namely full body, upper body with arms, torso and head, right arm and torso, left arm and torso, right arm alone, left arm alone, torso with legs, legs, right leg alone, and left leg alone. For each of these configurations we cluster the data as described above and learn poselet detectors. During test time we additionally run each detector for +/-7.5 degrees to compensate for slight rotations. Torso prediction is done using the detector from [16] that we augment with a spatial prior learned on the training set.

Next we present two different ways how the features f can be used to obtain image conditioned PS models.

3.2. Poselet Dependent Unary Terms

We first use the poselet features to obtain a location and rotation prediction for each body part separately.

Let us describe the location preference for a single part m only. During training, for part m, we cluster the relative distance between the torso and the part into $k = 1, \ldots, K$ clusters. For each cluster k we compute its mean offset from the torso μ_k and the variance of the differences Σ_k. This now forms a classification problem, from the poselet response f into the set of K clusters. To this end we train a classifier using sparse linear discriminant analysis (SLDA) [6] on the training set. We chose a sparse method since we expect a different set of poselets to be predictive for different body parts.

During test time we apply the learned classifier to predict from f the mean μ_k, and variance Σ_k that are subsequently used as a Gaussian unary potential for the part. We proceed analogously for rotation, that is we learn a classifier that predicts the absolute rotation of the body part based on poselet responses. Both unary parts together form a Gaussian potential $E^{u,poselet}$, and the complete set of unary terms of our model then reads

$$E^u(l_m; D) = E^{u,boost}(l_m; D) + w_p E^{u,poselet}(l_m; D), \tag{7}$$

where $E^{u,boost}$ is the original term given by Eq. 2 and w_p is the weighting parameter estimated on the validation set.

3.3. Poselet Dependent Pairwise Terms

To extend the pairwise terms we make them image dependent. For each pair of parts l_n, l_m we cluster their rela-

[1] The size of the region is set to 20×20 pixels in our experiments.

tive rotations into K clusters and obtain the parameters $\beta^{p,k}$ independently for each cluster using a maximum likelihood estimate. Similar to unary terms, we learn a SLDA classifier that predicts, given the feature f, into the set of clusters. This in turn yields the parameters β^p to be used for the image in question. The new pairwise potential that replaces E^p from Eq. 5 reads

$$E^{p,poselet}(l_n, l_m; D) = \langle \beta^p_{n,m}(f; D), \phi^p_{n,m}(l_n, l_m) \rangle. \quad (8)$$

We wrote $\beta(f)$ to make explicit its dependency on the poselet responses and that this parameter is being predicted.

4. Results

In this section we evaluate the proposed poselet-conditioned PS model on three well-known pose estimation benchmarks. We demonstrate that our new model achieves a significant improvement compared to the original PS model, while performing on par or even outperforming other competing approaches.

Datasets. For evaluation we use the following publicly available pose estimation benchmarks exhibiting strong variations in articulation and viewpoint: the recently proposed "Leeds Sports Poses" (LSP) dataset [14] that includes 1000 images for training and 1000 for testing showing people involved in various sports; the "Image Parsing" (IP) [17] dataset consisting of 100 train images and 205 test images of fully visible people performing various activities such as sports, dancing and acrobatics; the "UIUC People" dataset [23] consisting of 346 training and 247 test images of people in highly variable body poses playing different sports such as Frisbee or badminton. For each dataset we increase the training set size by adding the mirrored versions of the training images.

4.1. Results on LSP dataset

As in [14] we allocate 500 training images for the validation set and use it to estimate the weighting parameter in Eq. 7 and the number K of unary and pairwise clusters via grid search. The estimated values are $w_p = 0.05$ and $K = 12$. The poselets are trained as described in Section 3.1, which results in $P = 1036$ poselets. We follow [9] and use the observer-centric annotations provided by the authors of [9], which allows us to directly compare to their work. In the following we evaluate different model components and compare our approach to the best results in literature.

Using an oracle to select components. First we show the performance of our model assuming that the correct component for every potential is chosen by an oracle. This is the best case scenario that provides an upper bound on the performance our proposed model can achieve. We experimented with the number of components and found that 12 components per potential perform best. Increasing the number of components did not lead to improved results because of the limited number of training images available for parameter estimation for each component.

Results are shown in Tab. 1. It can be seen that adding poselet dependent terms improves the performance w.r.t. the baseline PS model [3]. Large improvements are consistently observed for all body parts. Correct predictions of unary rotation components improve the localisation of lower arms and legs most. This is explained by the fact that the rotation of these body parts is far less constrained compared to the rest of the limbs. Constraining part rotations to small ranges around the correct rotations reduces the uncertainty and steers the pose estimation towards the correct body pose. Similar effects can be seen when constraining positions of the unary potentials and learning the pairwise parameters from correct components, as this further constrains the predicted pose. The results show that using the parameters from correctly predicted components dramatically improves the localisation of all body parts in each particular setting. At the same time, the combination of all settings produces the best results which indicates that the constraints coming from different settings are complementary to each other. Note that even the model with oracle component prediction does not achieve values close to 100% because of test examples with extremely foreshortened or occluded body parts.

Evaluation of poselet-conditioned potentials. We evaluate each of the poselet-conditioned potentials described in Sec. 3 by plugging them one by one into our model. As each potential includes a classifier that maps poselet features to one of the components, we also evaluate the performance of these classifiers. The results are shown in Tab. 2. It can be seen that using PS + torso prediction improves the results compared to PS alone (56.2% vs. 55.7% PCP). Interestingly, when predicting the unary position parameters even despite the somewhat low component prediction accuracy of 43.9% we are able to improve the pose estimation result from 56.2% to 59.3% PCP. Similar results are obtained when predicting the unary rotation parameters (60.3% PCP). Combination of both further improves the performance to 60.8% PCP, as both potentials are complementary to each other.

We also analyse how prediction of pairwise parameters affects pose estimation. The prediction scores of pairwise components are generally lower than the absolute unary ones. A possible explanation is that the classification problem becomes harder because several rather different poselets might still correspond to the same relative angle between the two body parts. However, the final pose estima-

Setting	Torso	Upper leg	Lower leg	Upper arm	Forearm	Head	Total
Andriluka et al., [3]	80.9	67.1	60.7	46.5	26.4	74.9	55.7
+ predict unary rotation (ur)	96.4	91.1	86.1	76.6	60.2	88.5	81.3
+ predict unary position (up)	97.1	91.4	80.7	80.2	49.5	90.1	79.1
+ predict pairwise (p/wise)	93.2	88.5	81.6	73.6	58.0	87.6	78.4
+ ur + up + p/wise	**98.3**	**96.0**	**89.4**	**87.0**	**71.8**	**94.0**	**88.1**

Table 1. Pose estimation results (PCP) on the "Leeds Sport Poses" (LSP) dataset by our method *when using an oracle* to choose the correct component for every potential out of 12 possible values. This confirms the intuition that predicting the correct PS model directly translates to better PCP performance.

tion result is again improved (60.9% PCP). The combination of all three types of poselet-dependent potentials leads to further improvement and achieves 62.9% PCP. This indicates that the information provided by each type of potentials is complementary. Overall, our method achieves an improvement of 7.2% PCP over the original PS model that uses a generic pose prior. It shows that incorporating long range dependencies via mid-level feature representation can significantly boost the performance while keeping the inference efficient.

Comparison to the state of the art. We compare our method to competing approaches in Tab. 3. Interestingly, our method outperforms not only the baseline PS model (62.9% vs. 55.7% PCP), but also the state-of-the-art pose estimation model [26] which we downloaded from the authors' web page and retrained on the LSP dataset for fair comparison (62.9% vs. 60.8% PCP). The improvement is most prominent in case of localising upper legs (+6.2% PCP) whose configurations can be reliably captured by the legs- and torso-legs-poselets. The improvement is also pronounced for the lower legs which profit a lot from the improved upper legs localisation and for the upper arms (both +2.4% PCP). This result is very interesting since the method of [26] is a mixture of parts model that is quite different from ours, as it uses multiple unary templates for every part and image-independent pairwise potentials that do not allow to model long range part dependencies. In contrast, our model uses generic templates for each part, but incorporates a wide range of part unary terms by conditioning on poselet-representation. We also compare our method to the recent work [9], that extends the model of Yang&Ramanan by using additional background/foreground colour information across images of the same dataset and modify the hard negative mining procedure. Therefore, when comparing the numbers one has to bear in mind that the reported numbers of [9] are based on additional information about the dataset statistics. Compared to our method the difference is most pronounced in case of forearms where the skin colour information could be particularly helpful. Overall we conclude that both competing methods are orthogonal to our approach and are likely to improve when using multiple specific part templates and incorporating a colour model.

Setting	Avg. prediction accuracy, [%]	PCP, [%]
Andriluka et al., [3]	-	55.7
+ torso prediction	-	56.2
+ predict unary position (up)	43.0	59.3
+ predict unary rotation (ur)	37.4	60.3
+ ur + up	-	60.8
+ predict pairwise (p/wise)	30.8	60.9
+ up + ur + p/wise	-	**62.9**

Table 2. Accuracy of predicting a correct component for each unary and pairwise potential and corresponding pose estimation results (PCP) on the "Leeds Sport Poses" (LSP) dataset.

Method	Torso	Upper leg	Lower leg	Upper arm	Fore arm	Head	Total
ours	**87.5**	**75.7**	68.0	54.2	33.9	78.1	62.9
Andriluka et al., [3]	80.9	67.1	60.7	46.5	26.4	74.9	55.7
Yang&Ramanan [26]	84.1	69.5	65.6	52.5	35.9	77.1	60.8
Eichner& Ferrari [9]	86.2	74.3	**69.3**	**56.5**	**37.4**	**80.1**	**64.3**

Table 3. Pose estimation results on the "Leeds Sport Poses" (LSP) dataset with observer-centric annotations.

In Fig. 2 we show example pose estimation results using our method (row 1) and comparison to both [3] (row 2) and [26] (row 3). Our method is able to exploit long-range dependencies between parts accross a variety of activities such as tennis serve (columns 1 and 2), climbing (column 3) and running (column 4). In Fig. 3 (top row) we also show several examples of failure cases. The failure cases often correspond to images of people in poses that are underrepresented in the training set, and for which the prediction of unary and pairwise components is not accurate enough.

4.2. Results on IP dataset.

We now show the performance of our method on the "Image Parse" (IP) dataset. For evaluation we reuse the model learned on the LSP train set, but estimate the parameters w_p and K on the training set of the IP dataset. The

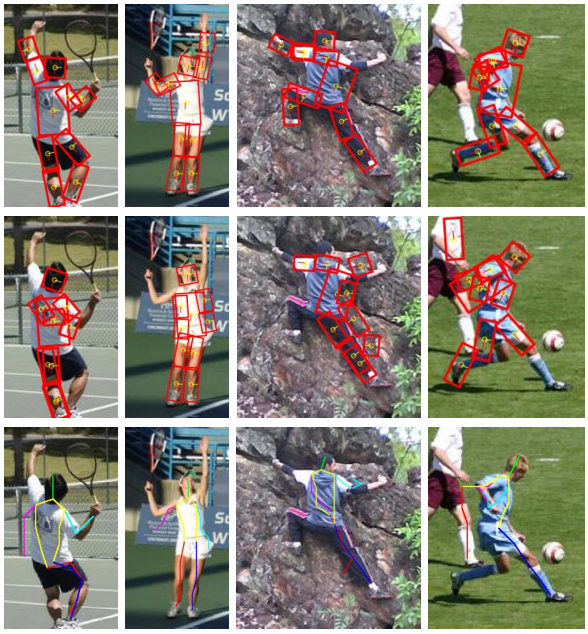

Figure 2. Sample pose estimation results on the LSP dataset obtained by our method (row 1), PS [3] and the method of [26] (row 3). Modelling long-range part dependencies by our method results in better performance on highly articulated people.

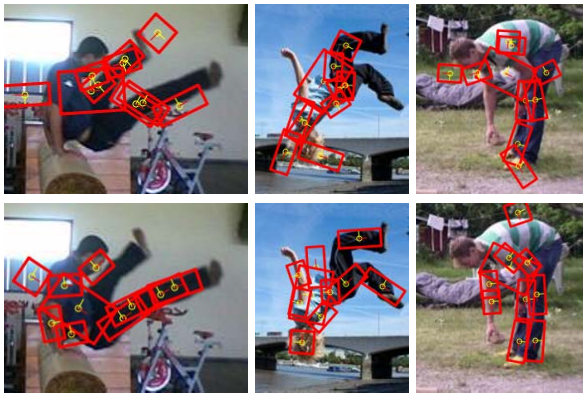

Figure 3. Typical failure cases on the LSP dataset. Shown are the results by our method (row 1) and PS [3] (row 2).

estimated values are $w_p = 0.1$ and $K = 12$. Note that the value of w_p increased with respect to the LSP dataset, which results in a stronger influence of the poselet features on the final solution. This could be due to a larger variability of people poses on the LSP dataset compared to IP (see [15] for the discussion and comparison of the two datasets).

The results are shown in Tab. 4. It can be seen that our method outperforms the baseline PS model (62.9% vs. 59.2% PCP), which is in line with the results on the larger LSP dataset. Our approach favourably compares to [26], outperforming it on all body parts apart from the lower arms. The most prominent improvement is observed for the torso, but the improvement for upper/lower legs is also pronounced. Our method is slightly better than the multi-layer

Method	Torso	Upper leg	Lower leg	Upper arm	Fore arm	Head	Total
ours	**92.2**	74.6	63.7	54.9	39.8	70.7	62.9
ours + [16]	90.7	**80.0**	**70.0**	59.3	37.1	77.6	66.1
Andriluka et al. [3]	86.3	66.3	60.0	54.6	35.6	72.7	59.2
Yang&Ramanan, [26]	82.9	69.0	63.9	55.1	35.4	77.6	60.7
Duan et al., [8]	85.6	71.7	65.6	57.1	36.6	**80.4**	62.8
Pishchulin et al., [16]	88.8	77.3	67.1	53.7	36.1	73.7	63.1
Johnson& Everingham, [15]	87.6	74.7	67.1	**67.3**	**45.8**	76.8	**67.4**

Table 4. Pose estimation results (PCP) on "Image Parse" (IP).

composite model of [8]. Their approach aims to capture non-tree dependencies between the parts by decomposing the model into multiple layers and performing dual decomposition to cope with cycles in the part graph. In contrast to their method, which incorporates multiple layers directly into the inference procedure making it infeasible without relaxations, our method implicitly models long-range dependencies between the parts and allows exact and efficient inference.

Our approach performs slightly worse compared to our approach [16], where we extended the tree-structured pictorial structures model with additional repulsive factors between non-adjacent parts and a stronger torso detector. We extend the approach in this paper with the repulsive factors and employ the same two-stage inference procedure as in [16]. The results are shown in Tab. 4. The extended model corresponds to "ours + [16]" and achieves 66.1% PCP, improving over all other models in the literature trained on the LSP dataset. Our result is only slightly worse than the result of the model from [15] that was trained on a significantly larger training set of 10000 images.

4.3. Results on UIUC People dataset.

For complete evaluation of our method we finally present results on the "UIUC People" dataset. We reuse the setting from the LSP dataset. We cluster the data into 20 clusters, again preserving only those containing at least 10 examples and learn poselet detectors on both UIUC+LSP data. The results are shown in Tab. 5. It can be seen that using only dataset-specific poselets already improves the results over the baseline PS model. This finding is consistent for all three datasets, we always improved when using poselet conditioned features. Interestingly, our method performs better than the approach of [25] that also falls behind the baseline PS model. This method is based on hierarchical poselets which intend to capture the non-tree dependencies between the parts via multiple layers. Such a model structure inevitably introduces cycles and requires an approximate inference.

Method	Torso	Upper arm	Lower arm	Upper arm	Fore arm	Head	Total
ours	**91.5**	**66.8**	**54.7**	38.3	**23.9**	**85.0**	**54.4**
Andriluka et al. [3]	88.3	64.0	50.6	**42.3**	21.3	81.8	52.6
Wang et al., [25]	86.6	56.3	50.2	30.8	20.3	68.8	47.0

Table 5. Pose estimation results (PCP) on the "UIUC People".

5. Conclusion

Pose estimation is often addressed with pictorial structures (PS) models based on a tree-structured graph leading to efficient and exact inference. However, tree-structured models fail to capture important dependencies between non-connected body parts leading to estimation failures. This work proposes to capture such dependencies using poselets that serve as a mid-level representation that jointly encodes articulation of several body parts. We show how an existing PS model for human pose estimation can be improved using a poselet representation. The resulting model is as efficient as the original tree-structured PS model, and is at the same time capable of representing complex dependencies between multiple parts. Experimental results show that a better prediction of human body layout using poselets improves body part estimation. We observe a consistent improvement on all of the considered datasets.

We believe that the components of our model could be further improved. In particular, future work should explore more robust and versatile mid-level features and other methods to condition the model on the image observations. For example, in addition to poselets a variety of other cues based on image motion, disparity and foreground segmentation could be use to adapt the model to the image at hand. One of the important limitations our current mid-level representation is its dependence of the torso detector, which could be a bottleneck in cases when the torso is obstructed by other body parts or scene objects. In the future we plan to explore representations that are more local and in addition to torso rely on a variety of other anchor points to establish the spatial correspondence between poselets. In the future we also plan to attend the most problematic cases of the current approach that are (self-)occlusion of body parts and fore-shortening. Finally, we envision that image-conditioned models based on mid-level representations could have applications beyond pose estimation, for example in activity recognition or object class detection.

References

[1] A. Agarwal and B. Triggs. Recovering 3D human pose from monocular images. *PAMI'02*. 3

[2] M. Andriluka, S. Roth, and B. Schiele. Discriminative appearance models for pictorial structures. *IJCV'11*. 1, 3, 4

[3] M. Andriluka, S. Roth, and B. Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *CVPR*, 2009. 2, 3, 4, 5, 6, 7, 8

[4] L. Bourdev, S. Maji, T. Brox, and J. Malik. Detecting people using mutually consistent poselet activations. In *ECCV*, 2010. 1

[5] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3D human pose annotations. In *ICCV'09*. 4

[6] L. Clemmensen, T. Hastie, D. Witten, and B. Ersbll. Sparse discriminant analysis. *Technometrics*, 2011. 4

[7] C. Desai and D. Ramanan. Detecting actions, poses, and objects with relational phraselets. In *ECCV*, 2012. 2

[8] K. Duan, D. Batra, and D. Crandall. A multi-layer composite model for human pose estimation. In *In BMVC'12*. 7

[9] M. Eichner and V. Ferrari. Appearance sharing for collective human pose estimation. In *In ACCV'12*. 5, 6

[10] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *PAMI'10*. 1, 3

[11] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *IJCV'05*. 1, 3

[12] M. A. Fischler and R. A. Elschlager. The representation and matching of pictorial structures. *IEEE Trans. Comput'73*. 1

[13] C. Ionescu, F. Li, and C. Sminchisescu. Latent structured models for human pose estimation. In *ICCV'11*. 3

[14] S. Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *BMVC'10*. 1, 2, 5

[15] S. Johnson and M. Everingham. Learning Effective Human Pose Estimation from Inaccurate Annotation. In *CVPR'11*. 1, 7

[16] L. Pishchulin, A. Jain, M. Andriluka, T. Thormaehlen, and B. Schiele. Articulated people detection and pose estimation: Reshaping the future. In *CVPR*, 2012. 4, 7

[17] D. Ramanan. Learning to parse images of articulated objects. In *NIPS'06*. 5

[18] G. Rogez, J. Rihan, S. Ramalingam, C. Orrite, and P. H. Torr. Randomized trees for human pose detection. In *CVPR'08*. 3

[19] B. Sapp, C. Jordan, and B. Taskar. Adaptive pose priors for pictorial structures. In *CVPR'10*. 3

[20] B. Sapp, D. Weiss, and B. Taskar. Parsing human motion with stretchable models. In *CVPR*, 2011. 1

[21] M. Sun and S. Savarese. Articulated part-based model for joint object detection and pose estimation. In *ICCV'11*. 1, 2

[22] T.-P. Tian and S. Sclaroff. Fast globally optimal 2d human detection with loopy graph models. In *CVPR'10*. 1

[23] D. Tran and D. A. Forsyth. Improved human parsing with a full relational model. In *ECCV*, 2010. 1, 5

[24] R. Urtasun and T. Darrell. Local probabilistic regression for activity-independent human pose inference. In *ICCV'09*. 3

[25] Y. Wang, D. Tran, and Z. Liao. Learning hierarchical poselets for human parsing. In *CVPR'11*. 1, 2, 4, 7, 8

[26] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *CVPR'11*. 1, 2, 3, 4, 6, 7