# Strong Appearance and Expressive Spatial Models for Human Pose Estimation

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

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

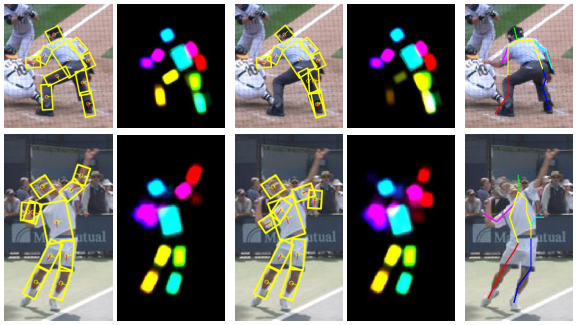
[2]Max Planck Institute for Intelligent Systems, Tübingen, Germany

## Abstract

*Typical approaches to articulated pose estimation combine spatial modelling of the human body with appearance modelling of body parts. This paper aims to push the state-of-the-art in articulated pose estimation in two ways. First we explore various types of appearance representations aiming to substantially improve the body part hypotheses. And second, we draw on and combine several recently proposed powerful ideas such as more flexible spatial models as well as image-conditioned spatial models. In a series of experiments we draw several important conclusions: (1) we show that the proposed appearance representations are complementary; (2) we demonstrate that even a basic tree-structure spatial human body model achieves state-of-the-art performance when augmented with the proper appearance representation; and (3) we show that the combination of the best performing appearance model with a flexible image-conditioned spatial model achieves the best result, significantly improving over the state of the art, on the "Leeds Sports Poses" and "Parse" benchmarks.*

## 1. Introduction

Most recent approaches to human pose estimation rely on the pictorial structures model representing the human body as a collection of rigid parts and a set of pairwise part dependencies. The appearance of the parts is often assumed to be mutually independent. Part detectors are either trained independently [17, 22] or jointly with the rest of the model [34, 7]. While effective detectors have been proposed for specific body parts with characteristic appearance such as heads and hands [20, 15], detectors for other body parts are typically weak. Obtaining strong detectors for all body parts is challenging for a number of reasons. The appearance of body parts changes significantly due to clothing, foreshortening and occlusion by other body parts. In addition, the spatial extent of the majority of the body parts is rather small, and when taken independently each of the parts lacks characteristic appearance features. For example lower legs often appear as a pair or parallel edges.



(a) (b) (c)

Figure 1. Example pose estimation results and corresponding part marginal maps obtained by (a) our full model combining local appearance and mid-level representation, (b) our best local appearance model and (c) results by Yang&Ramanan [34].

We argue that in order to obtain effective part detectors it is necessary to leverage both the pose specific appearance of body parts, and the joint appearance of part constellations. Pose specific person and body part detectors have appeared in various forms in the literature. For example, people tracking approaches [24, 14] rely on specialized detectors tailored to specific people poses that are easy to detect. Similarly, state-of-the-art approaches to people detection [3] build on a large collection of pose specific poselet detectors. Local [34] and global [17] mixture models that capture pose specific appearance of individual body parts and joints have shown to be effective for pose estimation. These approaches capture appearance at different levels of granularity: full person vs. subset of parts vs. individual parts and differ in the way they represent the appearance.

This paper builds on findings from the literature and follows two complementary routes to a more powerful pose model: improving the appearance representation and increasing the expressiveness of the joint body part model (see Fig. 1 and 3 for samples). Specifically, we consider local appearance representations based on rotation invariant or rotation specific appearance templates, mixtures of such local templates, specialized models tailored to appearance of

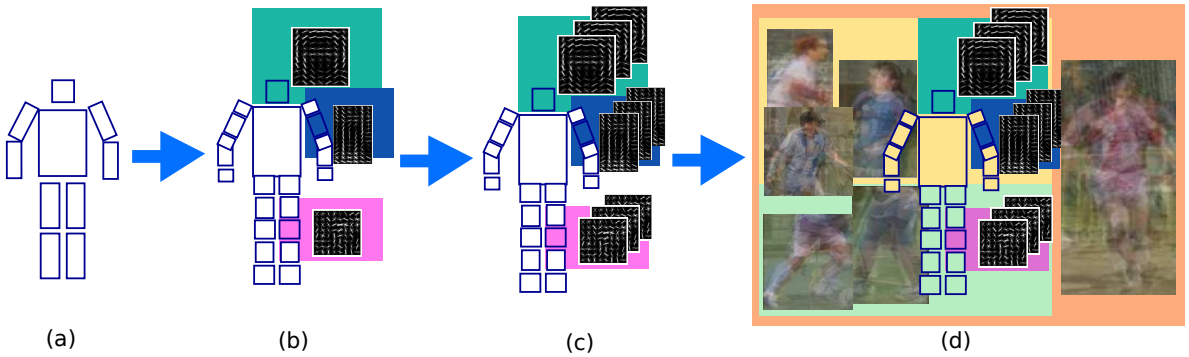



Figure 2. Overview of our method. We extend basic *PS* model [2] (a) to more flexible structure with stronger local appearance representations including single component part detectors (b) and mixtures of part detectors (c). Then we combine local appearance model with mid-level representation based on semi-global poselets which capture configurations of multiple parts (d). Shown are the means of sample poselet clusers. Color coding shows different levels of granularity of our appearance and spatial models.

salient body parts such as head and torso, and semi-global representations based on poselet features (Sec. 3). The second main contribution of the paper is to combine the improved appearance model with more expressive body representations. These include the flexible models of [26, 34] and the image conditioned spatial model of [21] (Sec. 4).

Starting with the basic tree-structured pictorial structures we perform a series of experiments incrementally adding various components and analysing the resulting performance gains (Fig. 2). Our analysis reveals several surprising facts (Sec. 5). The performance of the best appearance model for individual body parts is surprisingly high and can even compete with some approaches using weaker appearance terms but a full spatial model (Tab. 4). When augmented with the best appearance model, the basic tree-structured pictorial structures model perform superior to state-of-the-art models [9, 34] (Tab. 3). We show that strong appearance representations operating at different levels of granularity (mixtures of local templates vs. semi-global poselets) are complementary. Finally, we report the best result to date on the "Parse" and "Leeds Sports Poses" benchmarks, which are obtained by combining the best appearance model with the recently proposed image conditioned pictorial structures spatial model of [21] (Tabs. 5 and 6).

**Related work.** Various appearance representations have been considered in the past within the pictorial structures framework. Perhaps the most widespread representation is based on the discriminatively trained orientation invariant appearance templates [2, 26, 16, 23] composed of HOG [6] or shape context [19] features. These appearance models were extended by either including new types of features, or by generalising to mixtures of appearance templates. Several types of features have been proposed, such as skin and background color models [25, 9], part segmentation features [16, 32], image contours [25], pairwise color similarity [25, 29] and image motion features [26]. Interestingly, the best performing models today [16, 34] still build exclusively on the HOG feature, but do rely on the mixture appearance models. In our work we follow the best practices of the most successful models and build on traditional shape context and HOG-based features, while exploring different appearance representations.

Various local appearance models have been proposed, including stretchable models representing local appearance of body joints [34, 31, 26] and cardboard models modelling appearance of body parts as rigid templates [2, 25, 16]. Recently several works have been looking into semi-global representations based on multiple parts or poselets [15, 33] and global representations for entire bodies in various configurations [16, 22]. Also specialized models for detection of particular body parts, such as hands, head or entire upper body improve pose estimation results [4, 26, 12]. In this work we evaluate mentioned appearance representations and show that they are complementary to each other.

It was argued that for the tasks involving complex combinatorial optimisation strong detectors are especially important as they allow to effectively narrow down the search to the relevant part of the search space [30]. Pose estimation by detection has recently received more attention [20]. This requires models which in contrast to more traditional approaches focus on part detection and rely either on loose geometric features [26] or ignore them altogether [27, 15].

Other research is devoted to spatial modelling. Many methods use only one type of appearance and focus on other aspects such as efficient search [28, 25], or novel body models [34, 21] (discussed in Sec. 4). In this work we build on strong part detectors and demonstrate that even a basic tree-structure spatial human body model achieves state-of-the-art performance when augmented with the proper appearance representation. When combining strong appearance models with flexible image-conditioned spatial model, we outperform all current methods by a large margin.

## 2. Pictorial Structures Model

In the following we briefly summarise the basic tree-structured pictorial structures model [11, 13], that will serve as the baseline model for our analysis. In Sec. 3 and 4 we describe several extensions.

### 2.1. Model Formulation

The pictorial structures model represents the human body as a collection of rigid parts $L = \{l_1, \ldots, l_N\}$ and a set of pairwise part relationships. The state of each part is denoted by $l_n = (x_n, y_n, \theta_n, s_n)$, where $(x_n, y_n)$ is the image position of the part, $\theta_n$ is the absolute orientation, and $s_n$ is the part scale relative to the part size in the scale-normalised training set. Denoting the image observations by $D$, the energy of the body part configuration $L$ defined by the pictorial structures model is given by

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

The pairwise relationships between body parts are denoted by $n \sim m$. They follow the kinematic chain and thus result in a tree structured model.

We use the pictorial structures model introduced in [11] as our baseline model, and refer to it as *PS* in the remainder. This model is composed of $N = 10$ body parts: head, torso, and left and right upper arms, forearms, upper legs and lower legs. The parts are pairwise connected to form a tree corresponding to the kinematic chain, see Fig. 2(a). The pairwise terms $E^p$ encode the kinematic dependencies and are represented with Gaussians in the transformed space of joints between parts. We refer to the original paper [11] for the details on the pairwise terms. Note that in the basic model the spatial extent of each part, and in particular the distance between part centre and position of its joints is fixed, which potentially restricts the model to the configurations with relatively little foreshortening.

### 2.2. Learning and Inference

In this paper we use the publicly available implementation of the pictorial structures approach [1]. The parameters of unary and pairwise factors are learned using piecewise training. The pairwise term is set using a Maximum-Likelihood estimate that is available in closed form. The unary terms are described in Sec. 3.

Inference in the model is performed with sum-product belief propagation. Due to the tree structure this is an exact inference procedure yielding the marginal distributions for each body part. Predictions are then obtained by taking the maximum marginal state for each part. Some PS model variants that we will describe include auxiliary (latent) variables, this procedure thus marginalizes them out.

## 3. Better Appearance Representations

We now turn our attention to improving the appearance representations for body parts. These correspond to the unary terms $E^u$ in Eq. 1.

As the baseline model we consider the appearance representation introduced in [1]. These factors use boosted part detectors over shape context features, one detector per body part. This appearance representation is made independent to the part rotation, by normalising the training examples with respect to part rotation prior to learning. At test time, the detector is evaluated for each of 48 rotations in the discretized state-space of the PS model. The model that uses only this unary factor will be denoted as *PS*. [1]

### 3.1. Body Part Detectors

The rotation independent representation from [1] is based on a simplifying assumption, namely that the appearance of model parts does not change with part rotation. This typically is not true. For example the upper arms raised above the head and the ones held in front of the torso look quite different because of the overlap with other parts and change in the contours of the shoulders. This motivated rotation dependent detectors as in [34, 15].

We augment *PS* with two types of such local representations: 1) a rotation dependent detector tailored to the absolute orientation of the part (*rot-dep mix*) and 2) a rotation invariant representation tailored to a particular body pose (*pose-dep mix*). As an implementation we choose the deformable part model (DPM) [10] that has proven to be very reliable for detection purposes.

**Absolute Rotation.** Rotation dependent part detectors are obtained in the following way. We discretize the rotation space in $N = 16$ different bins, corresponding to a span of 22.5 degrees. All training data is assigned to the corresponding rotation bin based on the annotation. We then train a 16 component model, one component for each bin. As these models do capture rotation dependent appearance changes, we refer to this variant as *rot-dep mix*. A simpler baseline is a single component model trained for all rotations together. We include this model in the comparison under the name *rot-dep single*.

**Relative Rotation.** Rotation of the body parts is related to the orientation of the entire body, not necessarily to the absolute value in the image plane. We model this using a part detector that depends on the body pose. For this we normalise the part to a common rotation but rotate the entire body along with it. Then a binning in again 16 clusters is obtained by using the visibility features proposed in [7]. This clustering results in components that are compact w.r.t. the body pose in the proximity of the body part. The

[1] Please see [1] for further implementation details.

resulting detector is referred as *pose-dep mix*. Since this is "rotation invariant", in the sense the absolute rotation is irrelevant, during test time we evaluate this detector for all rotations in the state space of *PS*. We also include a simpler baseline which is a single component model trained from rotation-normalised body parts and then again evaluated for all rotations. We refer to it as *rot-inv single*.

### 3.2. Head and Torso Detectors (*spec-head*, *spec-torso*)

We consider two types of specialized part detectors proposed in the literature. The torso detector from [22] and the head detector from [18]. The main rationale behind using such specialized detectors is that body parts such as head and torso have rather specific appearance that calls for specialized part models.

Specifically, the torso detector of [22] is directly adapted from the articulated person detector based on a DPM. A torso prediction is obtained by regression using the positions of the latent DPM parts as features. This specialized torso detector benefits from evidence from the entire person and captures the pose. This is in contrast to the previous local torso model as it is not bound to evidence within the torso bounding box only. We refer to the specialized torso detector as *spec-torso*.

The head detector of [18] uses the observation that the main source of variability for the head is due to the viewpoint of the head w.r.t. the camera, *e.g.* front and profile views have a different but rather distinctive appearance. Following [18] we train a DPM detector for the head with $8$ components corresponding to a set of viewpoints discretized with a step of $45$ degrees. Note that the particular set of components is not available for the local detectors of the head that are either grouped by the in plane rotation or by the pose of the surrounding parts. We refer to specialized head detector as *spec-head*.

### 3.3. Implementation Details

All detectors outlined above are based on the DPM v4.0 framework and we utilise the publicly available software [10]. To turn a set of DPM detections after non-maximum suppression into a dense score for every pixel we apply a kernel density estimate (KDE). From the set $\{(d_k, s_k)\}, k = 1, \ldots, K$ with $d_k$ denoting the detection position and $s_k$ the detection score we define the score for part $l_n$ as the value of the KDE $E^u(l_n; D) = \log \sum_k w_k \exp(-\|l_k - d_k\|^2/\sigma^2)$, where $w_k = s_k + m$, and $m$ is a minimal detection score produced by the detector, which is set to $-3.0$ in our experiments. We then add the normalised DPM scores to the boosted part detector [2] scores at every position of the dense scoregrid and use these summed scores in the inference.

## 4. More flexible Models

Besides improving the pure appearance representations several works suggested to alter the model representation to make it more flexible. We incorporate their findings and include two modifications to the standard PS model.

### 4.1. Body Joints (*PS-flex*)

The original PS model represents body parts as variables, which in turn make appearance changes such as foreshortening very drastic. Follow-up work has suggested to build appearance representation for more local parts while allowing more flexibility in their composition [26, 34]. We incorporate this by including an additional 12 variables that represent location of the joints in the human body. These parts correspond to the left and right shoulder, elbow, wrist, hip, knee and ankle. In order to retain deterministic inference we incorporate these parts such that the resulting model is still tree-structured, as illustrated in Fig. 2(b). The additional pairwise terms between joint parts and body parts are modelled as a Gaussian factor w.r.t. their position. Since some body and joint parts are restricted to have the same absolute rotation, such as lower arm and wrist, we add a constraint on their rotation and scale to be identical. We refer to our flexible model as *PS-flex*.

### 4.2. Mid-level Representations (*mid-level*)

**Poselet Conditioned Deformation Terms.** The basic *PS* model has a limitation that the spatial distribution of the body parts is modelled as a Gaussian and can not properly represent the multi-modalities of human poses. We therefore take advantage of another extension from the literature [21] and substitute the unimodal image independent spatial factors in Eq. 1 with image conditioned factors. We define multiple pairwise terms for each joint by clustering the training data w.r.t. relative part rotation, and then predict the type of the pairwise term at test time based on the image features. To do so we train part configuration detectors called poselets and then use their responses during test time as mid-level feature representation (c.f. Fig. 2(d)). Prediction is treated as a multi-class classification problem where we use a classifier based on sparse linear discriminant analysis (sLDA) [5]. We denote this image conditioned flexible configuration as *mid-level p/wise*.

**Poselet Conditioned Appearance.** The local appearance models introduced in Sec. 3 are designed to capture pose dependent appearance of individual parts and pairs of adjacent parts. In order to capture appearance of the person at a higher level of granularity we extend our model with a mid-level poselet based representation and use poselet features described above to obtain rotation and position prediction of each body part separately. For instance, to predict part positions, we cluster the training data for each part based

on part relative offset w.r.t. torso centre into set of clusters. Then for each cluster its mean offset from the torso and the variance are computed. We then train a sLDA classifier to predict from the poselet features the mean and variance of the relative offset for every part and use these values as a Gaussian unary potential, which we add to other unary potentials introduced in Sec. 3. Prediction of absolute part orientation is done in a similar way. We call these representations in the following experiments as *mid-level rot* and *mid-level pos*, respectively and refer to [21] for further details on the implementation of these terms.

## 5. Results

In this section we evaluate the proposed extensions on two well-known pose estimation benchmarks and compare to other approaches from the literature. As a performance measure we use the common PCP loss [12].

**Datasets.** For evaluation we use the publicly available pose estimation benchmarks exhibiting strong variations in articulation and viewpoint: "Leeds Sports Poses" (LSP) dataset [16] that includes 1000 images for training and 1000 for testing showing people involved in various sports; the "Image Parsing" (IP) [23] dataset consisting of 100 train images and 205 test images of fully visible people in diverse set of activities such as sports, dancing and acrobatics.

### 5.1. Results on LSP dataset

In this section we report on the results obtained using the various extensions outlined in the last two sections. We follow [9] and use *observer-centric (OC)* annotations provided by the authors for evaluation. We train all the representations using the training set of LSP dataset.

**Flexible Model** We start with a comparison of models using body part appearance alone (*PS*) with the flexible model *PS-flex* that includes both joint and body part appearance. The results are shown in Tab. 1. We observe an improvement (+2.4%) due to better localization of lower legs and arms. This reinforces the findings of [34]: a flexible model of joints copes better with foreshortening. When removing the body parts for arms and legs and use only body joints (joints only) the performance drops. We attribute this to the easier confusion of joint detectors to background clutter. We conclude that the *PS-flex* model should benefit from better appearance representations which we will evaluate next.

**Single component detectors.** Performance of rotation dependent (*rot-dep single*) and rotation invariant (*rot-inv single*) single component detectors is reported in Tab. 2. Surprisingly, adding *rot-dep single* already improves the overall result (+2.7%), mostly due to better head localisation (+8.1%). The majority of the poses in the dataset are upright, thus much of head appearance change is captured by the *rot-dep single* detector. As expected, the result is further improved by *rot-inv single*, and the improvement is most prominent for lower arms (+7.8%). This clearly shows that rotation invariance of a single component detector is key to cope with the high degree of articulation by training and testing samples.

| Setting | Torso | Upper leg | Lower leg | Upper arm | Fore-arm | Head | Total |
|---|---|---|---|---|---|---|---|
| PS [2] | **80.9** | 67.1 | 60.7 | 46.5 | 26.4 | **74.9** | 55.7 |
| PS-flex (joints only) | 80.1 | 69.0 | 64.7 | 43.6 | 27.3 | 70.5 | 56.0 |
| PS-flex | 80.5 | **70.2** | **66.5** | **46.7** | **32.0** | 70.2 | **58.1** |

Table 1. Results on LSP when varying number of parts in PS.

**Mixtures of part detectors.** Rotation dependent mixture of detectors (*rot-dep mix*) accounts for the characteristic appearance changes of body parts under rotation. These types of detectors indeed improve the results, see line 4 in Tab. 2. When compared with the single counterparts we observe significant performance gain for all body parts.

While the former detectors are (in)variant to local rotations, they do not take the pose-specific appearance into account. The detectors *pose-dep mix* do. However, we do not observe any performance increase over *rot-dep mix*. We believe this is due to more compact cluster representations of the *rot-dep mix*, which makes them more discriminative. In summary, the best local mixture appearance representation improves over best single component detector by 2.8%, improving results for all parts. This indicates that mixtures better handle the highly multi-modal local appearance of body parts.

**Specialized detectors.** We discussed the possibility for designing specialized body part detectors in Section 3.2. We add those detectors to the *pose-dep mix* model, also including a Gaussian term on the torso location estimated via Maximum Likelihood on the training annotations. The results can be found in the last two lines of Tab. 2. Both the specialized torso and head detector improve the performance of torso and head localization, and via the connected model also improve the performance of other body parts. Even though the better torso prediction improves head localization (+0.3%), a specialized head detector still improves the performance (+1.1%). Since the parts are connected to the head via the torso, the influence of the *spec-head* detector on other body parts is found to be smaller. In summary, specialized detectors improve estimation results for all body parts, and give a +0.9% better results in terms of PCP. We expect this result would carry over to other models from the literature.

**Mid-level representations.** Now we combine the best performing local appearance representation with the mid-level representation of [21]. We use the same parameters as reported by the authors. Results are shown in Tab. 3.

| Setting | Torso | Upper leg | Lower leg | Upper arm | Fore- arm | Head | Total |
|---|---|---|---|---|---|---|---|
| PS-flex | 80.5 | 70.2 | 66.5 | 46.7 | 32.0 | 70.2 | 58.1 |
| + rot-dep single | 82.2 | 72.5 | 67.9 | 51.6 | 31.6 | 78.3 | 60.8 |
| + rot-inv single | 83.6 | 73.6 | 69.8 | 52.4 | 39.4 | 78.1 | 63.2 |
| + rot-dep mix | 87.2 | 76.0 | 72.2 | 55.9 | 40.5 | 83.3 | 66.0 |
| + pose-dep mix | 84.5 | 75.4 | 70.3 | 53.4 | 40.5 | 78.0 | 64.2 |
| + spec torso | 88.4 | 76.5 | 72.6 | 56.5 | 41.1 | 83.6 | 66.6 |
| + spec head | **89.2** | **76.7** | **72.8** | **56.9** | **41.2** | **84.7** | **66.9** |

Table 2. Results on LSP using local appearance models.

| Setting | Torso | Upper leg | Lower leg | Upper arm | Fore- arm | Head | Total |
|---|---|---|---|---|---|---|---|
| local appearance | 89.2 | 76.7 | 72.8 | 56.9 | 41.2 | 84.7 | 66.9 |
| + mid-level rot | 89.0 | 77.6 | 73.2 | 58.1 | 42.5 | 85.3 | 67.7 |
| + pos | **89.4** | 78.7 | **74.0** | 59.7 | 43.9 | **86.0** | 68.8 |
| + p/wise | 88.7 | **78.8** | 73.4 | **61.5** | **44.9** | 85.6 | **69.2** |

Table 3. Results on LSP using mid-level representations.

Predicting absolute orientation of parts based on mid-level representation (*mid-level rot*) noticeably improves results (+1.2%). Consistent improvement is achieved for each limb with forearms improving the most (+1.3%). Adding prediction of part positions based on mid-level features (*mid-level pos*) leads to further improvements (+1.1%). Again upper/lower arms profit the most from semi-global poselet detectors. They exhibit higher degree of articulation compared to other parts and thus are more difficult to detect using local detectors. Finally, adding prediction of pairwise terms (*mid-level p/wise*) improves the total performance, achieving an outstanding 69.2%. Overall, adding mid-level representations to the best performing local appearance model improves the results by 2.3%, giving improved results for all body parts. These results demonstrate the complementary effect of local appearance models and mid-level representations. Mid-level representation based on semi-global poselets models long range part dependencies, while local appearance model concentrate on local changes in the appearance of body parts.

**Performance using unaries only.** Finally we evaluate how much the appearance representation alone contributes to the final performance. To do so we remove all connections between the parts and evaluate part detectors only. Results are shown in Tab 4. As expected, boosted detectors of *PS-flex* perform worst. Adding our best local appearance model significantly improves the results (+13.6%), which demonstrates the strengths of the local appearance models compared to the original boosted detectors. Local mixtures of part detectors allow to model pose-dependent appearance of limbs while strong specific head and torso detectors push the performance of both most salient body parts (67.1 vs 36.2% for torso and 63.0 vs. 40.2% for head). Including the mid-level representation significantly improves the result further (+23.1%). So, upper/lower arms which are difficult to detect by local detectors profit a lot from semi-global poselets (+28.3 and +16.3%). A similar trend can be observed for upper/lower legs. This again demonstrates the strengths of mid-level representation and its complementary w.r.t. the local appearance models.

| Setting | Torso | Upper leg | Lower leg | Upper arm | Fore- arm | Head | Total |
|---|---|---|---|---|---|---|---|
| PS-flex | 36.2 | 20.1 | 27.1 | 6.8 | 5.5 | 40.2 | 19.5 |
| + local appearance | 67.1 | 36.2 | 35.2 | 18.6 | 10.6 | 63.0 | 33.1 |
| + mid-level | **79.5** | **65.5** | **63.5** | **46.9** | **26.9** | **77.1** | **56.2** |

Table 4. Performance on LSP using part appearance only.

**Comparison to the state of the art.** We compare our approach to other models from the literature in Tab. 5. Interestingly, our full model including local appearance and mid-level representations outperforms not only the baseline *PS* [2] (69.2 vs 55.7%), but all other current methods by quite a margin, improving 4.9% over the next best performing method [9]. The results also improve over [21] (69.2 vs. 62.9%) who uses similar mid-level representations but have a more simplistic local appearance model based on [2]. This is consistent for all body parts: torso +1.7%, upper legs +3.1%, lower leg +5.4%, upper arm +7.3%, forearm +11.0%, head +7.5%. We found this result interesting, as it clearly shows how much performance gain can be achieved by improving local part appearance while preserving the mid-level representation. We also compare our method to state-of-the-art pose estimation model [34] which we downloaded from the authors' web page and retrained on LSP dataset for fair comparison. Interestingly, our local appearance model combined with basic Gaussian pairwise terms already outperforms their method (66.9% vs. 60.8%). This demonstrates the strengths of the proposed local appearance model based on mixtures of pose-dependent detectors and specific torso and head detectors. When using our full model we outperform [34] by 8.4%. Finally, we compare our method to recent work [9], that extends the model [34] using additional background/foreground colour information across images of the same dataset and modify the hard negative mining procedure. Thus when comparing to [9] one should bear in mind that the reported numbers are based on additional information about the dataset statistics. Again, our local appearance model already performs better (66.9 vs. 64.3%). Comparing our full model, we observe an improvement of striking 4.9% over the current best result on LSP. This demonstrates the strength of combining local appearance modelling with flexible mid-level representations.

| Setting | Torso | Upper leg | Lower leg | Upper arm | Fore-arm | Head | Total |
|---|---|---|---|---|---|---|---|
| Our local appearance | **89.2** | 76.7 | 72.8 | 56.9 | 41.2 | 84.7 | 66.9 |
| Our full model | 88.7 | **78.8** | **73.4** | **61.5** | **44.9** | **85.6** | **69.2** |
| Andriluka et al., [2] | 80.9 | 67.1 | 60.7 | 46.5 | 26.4 | 74.9 | 55.7 |
| Yang&Ramanan [34] | 84.1 | 69.5 | 65.6 | 52.5 | 35.9 | 77.1 | 60.8 |
| Pishchulin et al., [21] | 87.5 | 75.7 | 68.0 | 54.2 | 33.9 | 78.1 | 62.9 |
| Eichner&Ferrari [9] | 86.2 | 74.3 | 69.3 | 56.5 | 37.4 | 80.1 | 64.3 |

Table 5. Comparison of pose estimation results (PCP) on LSP dataset to current methods using observer-centric (OC) annotations. Results using person-centric (PC) annotations available here `www.d2.mpi-inf.mpg.de/poselet-conditioned-ps`

**Qualitative evaluation.** Successful results of our model are shown in Fig. 3 (rows 1-4). Our local appearance model already achieves good results (Fig. 3(b)), as it is able to cope with highly variable part appearance. Our full model which also includes mid-level representations further improves the results (Fig. 3(a)), as it captures the entire pose of the body and models other part dependencies. This is in contrast to Yang&Ramanan [34] (Fig. 3(c)) who rely only on local image evidence. Typical failure cases of our model include large variations in scale between body parts (Fig. 3 (line 5)), untypical appearance and poses (line 6) and massive self-occlusion (line 7).

### 5.2. Results on Image Parse dataset

In the experiments on the Image Parse dataset [23] we use our full model trained on the LSP dataset and set the parameters of the mid-level representation as reported by [21]. In Tab. 6 we compare our full model with a number of recent approaches from the literature. Our method improves over the next best performing method by 2.0%.

We outperform the recent work [21] (+6.5%). Their model uses a similar mid-level representation but their local appearance is based on *PS* [2]. This result is in line with the findings on LSP, and shows the importance of better appearance models. Our method consistently improves over the pose estimation model of Yang&Ramanan [34] (+8.7%) and the over the newer version of this model from [35] (+2.3%). The improvement is achieved for all body parts apart from head and lower legs. In particular, we improve on highly articulated forearms (+6.1%) and upper legs (+2.2%). This demonstrates that much improvement can be gained from the complementary mid-level representation. Our result is also significantly better than the multi-layer composite model of [8] (+6.6%), who captures non-tree part dependencies by decomposing the model into several layers and using dual decomposition to cope with the resulting loopy graph. In contrast, our method implicitly models long-range dependencies between the parts by using mid-level representation while allowing exact and efficient inference. We outperform the method of [22] (+6.3%), who also integrate

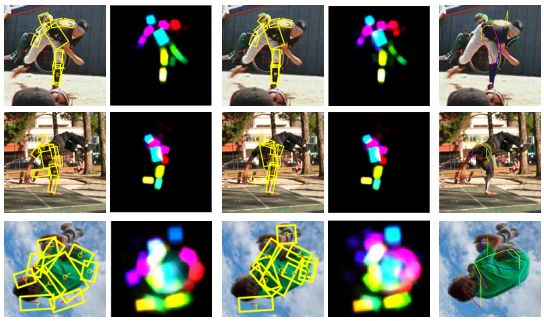




Figure 3. Qualitative results: estimated poses and corresponding part marginal maps obtained by (a) our full model combining local appearance and flexible mid-level representation, (b) our local appearance model and (c) results by Yang&Ramanan [34].

the evidence from a people detector into the PS framework to improve torso localisation. Their method introduces loops between the corresponding upper/lower legs to prevent over-counting, again yielding more expensive inference. Finally, our method outperforms [17] (+2.0%), the best published result on this dataset. Note that their model also uses strong local appearance models and is trained on an additional dataset of 10000 images.

## 6. Conclusion

In this paper we investigated the use of 1) stronger appearance models and 2) more flexible spatial models. We observe that better local appearance representations directly result in better performance and even a basic tree-structured human body model achieves state-of-the-art performance when augmented with the proper appearance representa-

| Setting | Torso | Upper leg | Lower leg | Upper arm | Fore-arm | Head | Total |
|---|---|---|---|---|---|---|---|
| Our full model | **93.2** | 77.1 | 68.0 | 63.4 | **48.8** | 86.3 | **69.4** |
| Andriluka et al. [1] | 86.3 | 66.3 | 60.0 | 54.6 | 35.6 | 72.7 | 59.2 |
| Yang&Ramanan [34] | 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., [22] | 88.8 | **77.3** | 67.1 | 53.7 | 36.1 | 73.7 | 63.1 |
| Pishchulin et al., [21] | 92.2 | 74.6 | 63.7 | 54.9 | 39.8 | 70.7 | 62.9 |
| Yang&Ramanan [35] | 85.9 | 74.9 | **68.3** | 63.4 | 42.7 | **86.8** | 67.1 |
| Johnson& Everingham, [17] | 87.6 | 74.7 | 67.1 | **67.3** | 45.8 | 76.8 | 67.4 |

Table 6. Comparison of pose estimation results (PCP) on "Image Parse" dataset to current methods.

tion. The second route explored in this paper are more flexible spatial body models with image conditioned terms based on mid-level representations, implemented as poselets. We find significant improvement using this information, both when using a connected and even a disconnected body model. The effects of the terms studied are found to be additive, the combination significantly outperforms all competitors as demonstrated on two benchmark datasets. The source code of our approach will be made publicly available [2]. Note that all representations considered in this paper rely on the image gradient information only. We will aim at incorporating other image features in the future.

## References


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

[2] M. Andriluka, S. Roth, and B. Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *CVPR'09*.

[3] L. Bourdev, S. Maji, T. Brox, and J. Malik. Detecting people using mutually consistent poselet activations. In *ECCV'10*.

[4] P. Buehler, M. Everingham, D. P. Huttenlocher, and A. Zisserman. Long term arm and hand tracking for continuous sign language TV broadcasts. In *BMVC'08*.

[5] L. Clemmensen, H. Trevor, W. Daniela, and E. Bjarne. Sparse discriminant analysis. *Technometrics'11*.

[6] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR'05*.

[7] C. Desai and D. Ramanan. Detecting actions, poses, and objects with relational phraselets. In *ECCV'12*.

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

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

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

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

[12] V. Ferrari, M. Marin, and A. Zisserman. Progressive search space reduction for human pose estimation. In *CVPR'08*.

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

[14] A. Fossati, M. Dimitrijevic, V. Lepetit, and P. Fua. Bridging the gap between detection and tracking for 3D monocular video-based motion capture. In *CVPR'07*.

[15] G. Gkioxari, P. Arbelaez, L. Bourdev, and J. Malik. Articulated pose estimation using discriminative armlet classifiers. In *CVPR'13*.

[16] S. Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *BMVC'10*.

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

[18] M. Marin-Jimenez, A. Zisserman, and V. Ferrari. "here's looking at you, kid." detecting people looking at each other in videos. In *In BMVC'11*.

[19] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *PAMI'05*.

[20] A. Mittal, M. Blaschko, A. Zisserman, and P. Torr. Taxonomic multi-class prediction and person layout using efficient structured ranking. In *ECCV'12*.

[21] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele. Poselet conditioned pictorial structures. In *CVPR'13*.

[22] L. Pishchulin, A. Jain, M. Andriluka, T. Thormaehlen, and B. Schiele. Articulated people detection and pose estimation: Reshaping the future. In *CVPR'12*.

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

[24] D. Ramanan, D. A. Forsyth, and A. Zisserman. Strike a pose: Tracking people by finding stylized poses. In *CVPR'05*.

[25] B. Sapp, A. Toshev, and B. Taskar. Cascaded models for articulated pose estimation. In *ECCV'10*.

[26] B. Sapp, D. Weiss, and B. Taskar. Parsing human motion with stretchable models. In *CVPR'11*.

[27] 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 a single depth image. In *CVPR'11*.

[28] M. Sun, M. Telaprolu, H. Lee, and S. Savarese. An efficient branch-and-bound algorithm for optimal human pose estimation. In *CVPR'12*.

[29] D. Tran and D. A. Forsyth. Improved human parsing with a full relational model. In *ECCV'10*.

[30] Z. Tu, X. Chen, A. L. Yuille, and S. chun Zhu. Image parsing: Unifying segmentation, detection, and recognition. *IJCV'03*.

[31] N. Ukita. Articulated pose estimation with parts connectivity using discriminative local oriented contours. In *CVPR'12*.

[32] H. Wang and D. Koller. Multi-level inference by relaxed dual decomposition for human pose segmentation. In *CVPR'11*.

[33] Y. Wang, D. Tran, and Z. Liao. Learning hierarchical poselets for human parsing. In *CVPR'11*.

[34] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *CVPR'11*.

[35] Y. Yang and D. Ramanan. Articulated human detection with flexible mixtures of parts. *PAMI*, to appear.


[2] `www.d2.mpi-inf.mpg.de/poselet-conditioned-ps`