# Expressive Models and Comprehensive Benchmark for 2D Human Pose Estimation

Leonid Pishchulin[1], Mykhaylo Andriluka[1,3], Peter Gehler[2], and Bernt Schiele[1]

[1] Max Planck Institute for Informatics, Germany
[2] Max Planck Institute for Intelligent Systems, Germany
[3] Stanford University, USA

## 1 Introduction

In this work we consider the challenging task of articulated human pose estimation in monocular images. Most of current methods in this area [4, 8, 16, 14] 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.

In this work we advance the state of the art in articulated human pose estimation in three ways. *First*, we argue that modeling part dependencies between *non-adjacent* body parts is important for effective pose estimation (cf. Fig. 1). We propose a model [10] 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 and retains an exact and efficient inference procedure in a tree-based model. *Second*, we explore various types of appearance representations with the aim to improve the body part hypotheses [11]. 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. We show that the proposed appearance representations are complementary and a combination of the best performing appearance model paired with a flexible image-conditioned spatial model achieves the best result. *Third*, we introduce a novel benchmark "MPII Human Pose"[4] [3] that makes a significant advance in terms of diversity and difficulty, a contribution that we feel is required for future developments in human body models. This comprehensive dataset was collected using an established taxonomy of over 800 human activities. The collected images cover a wider variety of human activities than previous datasets including various recreational, occupational, and householding activities. People are captured from a wider range of viewpoints. In addition we provide a rich set of labels including positions of body joints, full 3D torso and head orientation, occlusion labels for joints and body parts, and activity labels. With these annotations we perform a detailed analysis [3, 12] of the leading 2D human pose estimation and activity recognition methods to understand success and failure cases for established models.

[4] Available at http://human-pose.mpi-inf.mpg.de.

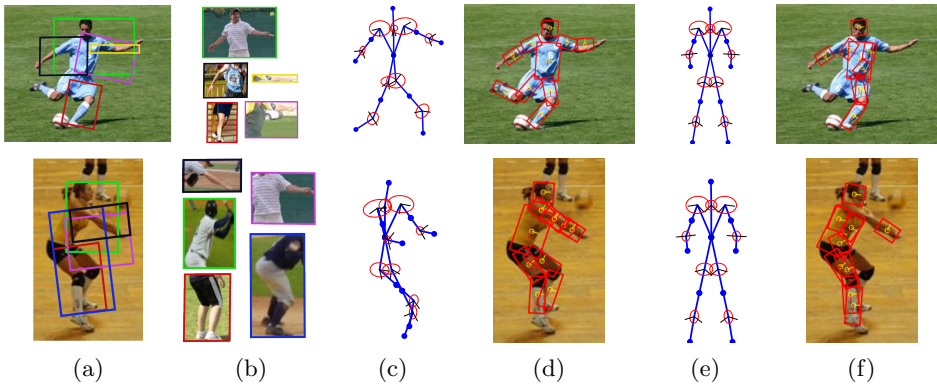


**Fig. 1.** Visualization of our approach. (a) shows the top scoring poselet detections with the corresponding poselet cluster medoids (b). It is visible that the poselets capture the anatomical configuration of the human in the input image. All poselet detections contribute to a prediction of the deformable 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 [2] (e) and the corresponding pose prediction (f).

## 2 Poselet Conditioned Pictorial Structures

The approach [10] is based on the question how mid-level representations of anatomical configurations of human poses can predict an image-specific pictorial structures (PS) model that in turn is applied to the image. This representation is inspired by the work [5, 15] 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.

This model is visualized in Fig. 1. From 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 predicted by our method. A pose of a person that was estimated with this poselet-conditioned model is shown in Fig. 1(d). For comparison we show the deformation model of [2] (a generic pose prior, the same for all images) along with the corresponding pose estimate in the last two columns.

***Deformation terms.*** 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 poselet detectors and then use their responses during test time as mid-level feature representation. Prediction is treated as a multi-class classification problem.


**Fig. 2.** Randomly chosen images from a set of activity categories of the proposed "MPII Human Pose" dataset. Image captions indicate activity category (1st row) and activity (2nd row). To view the full dataset visit http://human-pose.mpi-inf.mpg.de.

***Appearance terms.*** In order to capture appearance of the person at a higher level of granularity we 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 center. Then for each cluster its mean offset from the torso and the variance are computed. We then train a multi-class 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. Prediction of the absolute part orientation is done in a similar way.

## 3 Strong Appearance Representations

We now turn our attention to improving the local appearance representations for individual body parts and explore various types of appearance representations [11]. Below we describe the representations we found to perform best.

***Local part detectors.*** The appearance of individual body parts changes with part rotations and therefore we augment the model with rotation dependent part detectors. These are obtained in the following way. The rotation space is discretized in 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.

***Head and torso detectors.*** In addition to local part detectors, we consider two types of specialized detectors proposed in the literature: the torso detector adapted from the articulated person detector [13] based on a DPM [6], and the head detector inspired by [9]. 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.

| Setting | Torso | Upper leg | Lower leg | Upper arm | Fore-arm | Head | Upper body | Full body |
|---|---|---|---|---|---|---|---|---|
| our full model [10, 11] | **63.8** | **39.6** | **37.3** | **39.0** | **26.8** | **70.7** | **39.1** | **42.3** |
| Yang&Ramanan [16] | 61.0 | 36.6 | 36.5 | 34.8 | 17.4 | 70.2 | 33.1 | 38.3 |
| Gkioxari et al. [7] | 51.3 | - | - | 28.0 | 12.4 | - | 26.4 | - |
| Sapp&Taskar [14] | 51.3 | - | - | 27.4 | 16.3 | - | 27.8 | - |

**Table 1.** Pose estimation results (PCPm) on the proposed dataset.

## 4 Dataset

Current datasets are limited in their coverage of the challenges that are encountered in a general pose estimation setup. Still, they serve as the common sources to train and evaluate different models. In our recent work [3] we present a large dataset of images that covers a wide variety of human poses and clothing types of people interacting with various objects and environments. This dataset was collected from YouTube videos using an established two-level hierarchy of over 800 every day human activities [1]. The activities at the first level of the hierarchy correspond to thematic categories such as "Home repair", "Occupation", "Music playing", etc., while the activities at the second level correspond to individual activities, e.g., "Painting inside the house", "Hairstylist", and "Playing woodwind". In total, the dataset contains 20 categories and 410 individual activities covering a wide variety of different human activities. Due to the systematic coverage this dataset is representative of the diversity of human poses, overcoming one of the main limitations of previous collections. Overall the dataset consists of 25K images containing over 40K people with annotated body joints, from which we allocate roughly three quarters for training. In addition, for the test set we provide richer labels including full 3D torso and head orientation and occlusion labels for joints and body parts. These labels enable a thorough analysis [3, 12] of the factors leading to successes and failures of current pose estimation and activity recognition methods.

## 5 Results and Conclusion

We evaluate the performance of our model on the proposed "MPI Human Pose" dataset and compare to the results by published methods. We use the PCPm measure [3] for evaluation. It can be seen that the proposed method [10, 11] significantly outperforms the full body method by Yang&Ramanan [16], as well as the upper body methods by Gkioxari et al. [7] and Sapp&Taskar [14]. Our method is also among the top performing on standard pose estimation benchmarks. See [11] for more results.

Our analysis on the "MPI Human Pose" dataset indicates that current methods are challenged by large torso rotation and loose clothing. From all other factors, pose complexity has the most profound effect on the performance. Current methods perform best on activities with simple tight clothing (e.g. in sport scenes) and are challenged by images with complex clothing and background clutter that are typical for many occupational and outdoor activities.

## References


1. Ainsworth, B., Haskell, W., Herrmann, S., Meckes, N., Bassett, D., Tudor-Locke, C., Greer, J., Vezina, J., Whitt-Glover, M., Leon, A.: 2011 compendium of physical activities: a second update of codes and MET values. MSSE'11
2. Andriluka, M., Roth, S., Schiele, B.: Pictorial structures revisited: People detection and articulated pose estimation. In: CVPR'09
3. Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: 2d human pose estimation: New benchmark and state of the art analysis. In: CVPR'14
4. Andriluka, M., Roth, S., Schiele, B.: Discriminative appearance models for pictorial structures. IJCV'11 http://dx.doi.org/10.1007/s11263-011-0498-z
5. Bourdev, L., Malik, J.: Poselets: Body part detectors trained using 3D human pose annotations. In: ICCV'09
6. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. PAMI'10
7. Gkioxari, G., Arbelaez, P., Bourdev, L., Malik, J.: Articulated pose estimation using discriminative armlet classifiers. In: CVPR'13
8. Johnson, S., Everingham, M.: Learning Effective Human Pose Estimation from Inaccurate Annotation. In: CVPR'11
9. Marin-Jimenez, M., Zisserman, A., Ferrari, V.: "here's looking at you, kid." detecting people looking at each other in videos. In: In BMVC'11
10. Pishchulin, L., Andriluka, M., Gehler, P., Schiele, B.: Poselet conditioned pictorial structures. In: CVPR'13
11. Pishchulin, L., Andriluka, M., Gehler, P., Schiele, B.: Strong appearance and expressive spatial models for human pose estimation. In: ICCV'13
12. Pishchulin, L., Andriluka, M., Schiele, B.: Fine-grained activity recognition with holistic and pose based features. In: GCPR'14
13. Pishchulin, L., Jain, A., Andriluka, M., Thormaehlen, T., Schiele, B.: Articulated people detection and pose estimation: Reshaping the future. In: CVPR'12
14. Sapp, B., Taskar, B.: Multimodal decomposable models for human pose estimation. In: CVPR'13
15. Wang, Y., Tran, D., Liao, Z.: Learning hierarchical poselets for human parsing. In: CVPR'11
16. Yang, Y., Ramanan, D.: Articulated human detection with flexible mixtures of parts. PAMI'13