Articulated People Detection and Pose Estimation: Reshaping the Future

Leonid Pishchulin Arjun Jain Mykhaylo Andriluka
Thorsten Thormählen Bernt Schiele
Max Planck Institute for Informatics, Saarbrücken, Germany

Abstract

State-of-the-art methods for human detection and pose estimation require many training samples for best performance. While large, manually collected datasets exist, the captured variations w.r.t. appearance, shape and pose are often uncontrolled thus limiting the overall performance. In order to overcome this limitation we propose a new technique to extend an existing training set that allows to explicitly control pose and shape variations. For this we build on recent advances in computer graphics to generate samples with realistic appearance and background while modifying body shape and pose. We validate the effectiveness of our approach on the task of articulated human detection and *articulated pose estimation. We report close to state of the art results on the popular Image Parsing [25] human pose estimation benchmark and demonstrate superior performance for articulated human detection. In addition we define a new challenge of combined articulated human detection and pose estimation in real-world scenes.*

1. Introduction

Recent progress in people detection and articulated pose estimation may be contributed to two key factors. First, discriminative learning allows to learn powerful models on a large training corpora [6, 11, 31]. Second, robust image features enable to deal with image clutter, occlusions and appearance variation [7, 22]. Large and representative training sets are essential for best performance and significant effort has been made collecting them [6, 20, 10]. Typically, images are extracted from public data sources (e.g. photo collections) and manually annotated. However, even for large datasets it remains a challenge to ensure that they adequately cover the space of possible body poses, shapes and appearances. Even more importantly, depending on the task (e.g. detecting people in basketball vs. golf vs. street-scenes) the *relevant* distribution of shape, body pose and appearance varies greatly and cannot be easily controlled using manually collected datasets.

In this paper we are interested in the challenging problem of articulated people detection *and* pose estimation in challenging real-world scenes. In order to achieve this goal (e.g. illustrated in Fig. 1), we advance the state of the art in several ways. As a first contribution, we propose a novel method for automatic generation of multiple training examples from an arbitrary set of images with annotated human body poses. We use a 3D human shape model [16] to produce a set of realistic shape deformations of person's appearance, and combine them with motion capture data to produce a set of feasible pose changes. This allows us to generate realistically looking training images of people where we have full control over the shape and pose variations. As a second contribution, we evaluate our data generation method on the task of articulated human detection and on the task of human pose estimation. We explore how various parameters of the data generation process affect overall performance. On both tasks we can significantly improve performance when the training sets are extended with the automatically generated images. As a third contribution, we propose a joint model that directly integrates evidence from an appropriately trained deformable part model (DPM, [11]) into a pictorial structures framework and demonstrate that this joint model further improves performance. Last, as fourth contribution, we define a new challenge of joint detection and pose estimation of multiple articulated people in challenging real-world scenes.

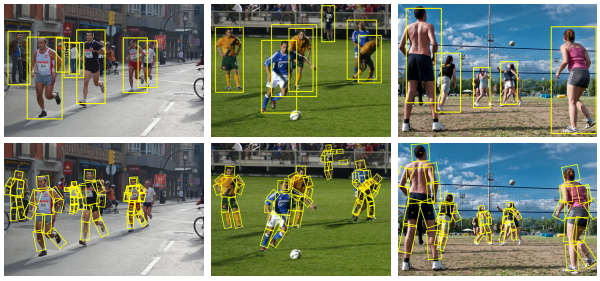


Figure 1: Sample detections (top) and pose estimates (bottom) of multiple articulated people obtained with our model trained on images from our new data generation method.

Related work. People detection and articulated pose estimation are closely related and challenging problems. Much recent work focuses on special cases such as detection of pedestrians [7, 9, 8] or people seen mostly from frontal

To appear in the Proc. of the *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Providence, Rhode Island, June 2012.

views [13, 27]. Approaches designed for generic people detection [6, 11] are often evaluated on the VOC [10] benchmark mostly capturing upright people. While there exist a significant body of literature addressing pose estimation in challenging real-world scenes [25, 30, 20], they typically assume that people were localized in the images. With a few notable exceptions [2, 17, 31] the task of joint detection and pose estimation of strongly articulated people remains largely unaddressed.

Training state-of-the-art models for detection of strongly articulated people requires representative training sets. Collecting and annotating such data sets is tedious and many images are required for good performance [20]. Here, we follow the appealing route to generate training data based on computer graphics methods. Automatically generated data has been used in computer vision in the past. However, its application has been mostly limited to cases where realistic appearance is not required, such as silhouette-based methods for human pose estimation [1] or depth images [29]. While training people detectors from rendered images has been proposed [23, 21, 28], such training data often lacks the necessary realism for good performance. An alternative is to apply transformations to real images preserving their realism. E.g. [9] augments the training set by applying a morphable 2D model to images. Here we follow a similar idea, however in our case we use a generative 3D human shape model and motion capture data to generate possible deformations of 2D data making our deformation model more realistic and versatile. [24, 32] are probably closest to our work. In our own prior work [24] we require an expensive data acquisition step limiting the number of subjects in the experiments to a handful of people. Both methods are limited to shape deformations *only*. On the contrary in this work we are able to generate new training examples from existing 2D images while still allowing for a wide range of shape *and* pose variations. We show that controlling pose variations of generated training samples is essential when training detection models of highly articulated people.

2. Generation of novel training examples

To improve both articulated people detection and pose estimation we aim to generate training images with full control over pose and shape variations. Fig. 2 gives an overview of our novel data generation process consisting of three stages. Starting from approximate 3D pose annotations we first recover the parameters of the 3D human shape model [16]. The body shape is then modified by *reshaping* and *animating*. Reshaping changes the shape parameters according to the learned generative 3D human shape model and animating changes the underlying body skeleton. Given the new reshaped and/or animated 3D body shape we back-project it into the image and morph the segmentation of the person. To that end we employ the linear blend skinning

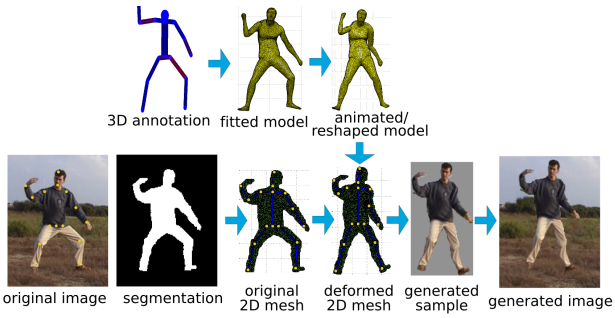


Figure 2: Overview of our novel data generation method.

procedure with bounded biharmonic weights described in [18]. The following describes these steps in more detail.

2.1. Data annotation

For each subject in the training set we manually provide a 3D pose and a semi-automatic segmentation of the person. The 3D pose is obtained using the annotation tool introduced in [6]. The pose is used later to resolve the depth ambiguities which otherwise arise when fitting the 3D human shape model to 2D observations. The initial segmentation is obtained with GrabCut [26] which we automatically initialize using annotated 2D joint positions and projected 3D shape from the fitted shape model (see below). While this procedure already produces reasonable results, segmented images often require user interaction to refine the segmentation due to low resolution, poor contrast and bad lighting. We use the segmentation to compute a 2D image mesh which is then deformed to change human shape and pose.

2.2. 3D human shape recovery and animation

3D human shape model. In order to generate photorealistic synthetic images of people in different poses we employ a statistical model of 3D human shape and pose [16] which is a variant of the SCAPE model [3]. The model is learned from a public database of 3D laser scans of humans and thus represents the available shape and pose variations in the population. The shape variation across individuals is expressed via principal component analysis (PCA). We use the first 20 PCA components capturing 97% of the body shape variation. Linear blend skinning is used to perform pose changes. To this end, a kinematic skeleton was rigged into the average human shape model. The 3D model pose is represented by a kinematic skeleton with 15 joints having a total of 24 degrees of freedom (DoF) plus 6 DoF for global body position and orientation. The model surface consists of a triangle mesh with 6450 vertices and 12894 faces.

Model fitting. Having an annotated 3D pose allows to resolve the depth ambiguity while fitting the 3D shape model's kinematic skeleton to a 2D image. We retarget the skeleton to an annotated 3D pose by computing inverse kinematics through minimizing the Euclidean distance be-

tween a set of corresponding 3D joint positions, namely left/right ankles, knees, hips, wrists and elbows, upper neck and head. We use a constrained optimization based on the iterative interior point method. Optimization is done in shape and pose parameters space. Obtaining a good fit of the skeleton is essential for the rest of our data generation process and can significantly influence the realism of generated images. The fitting dependents on the flexibility of the kinematic skeleton and also on how well the corresponding 3D joint positions match. We thus do not include shoulders, pelvis and thorax joints into the objective function as these tend to have different positions in the annotated 3D pose and the 3D model's kinematic skeleton.

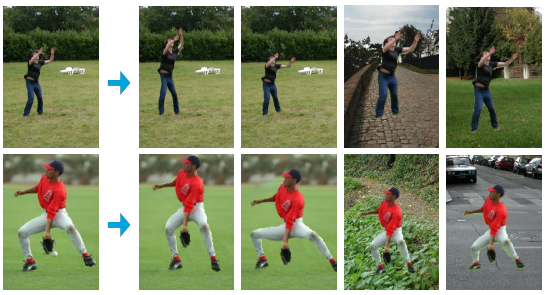
Varying model shape and pose. After fitting the skeleton we vary the 3D shape and pose parameters. To change the shape we randomly sample from the underlying 3D human shape distribution. For 3D shape animation we require a database of poses. To that end we retargeted the shape model's kinematic skeleton to over 280,000 of highly articulated poses from freely available mocap data[1]. To do so, we fix the bone lengths of the mocap skeleton to be the same as for the shape model's skeleton and compute inverse kinematics by optimizing over global rotation, translation and pose parameters only, which reduces the search space and produces better results. To animate the fitted skeleton we use the nearest retargeted poses with an average joint distance of less than 90 mm. Informal experiments showed that going further away from the fitted pose may result in unrealistically looking generated images.

2.3. Generation of novel images

After shape and pose changes are applied to the fitted 3D shape model, we project its 3D joint positions into the image and move 2D annotated joints towards corresponding projected joints. This results in a smooth 2D mesh deformations described by linear blend skinning [18]. We only animate "dangling" arms and legs, and do not deform occluded or occluding limbs as this leads to unrealistic deformations.

To obtain a final training sample we render the deformed 2D mesh into a photorealistically looking individual by reusing the original appearance of the person. Finally we combine the rendered subject with the background. We either replace the original person with the generated one by first removing the original person from the image using a commercial implementation of [4], or embed the generated sample at a random place of a new people-free image. Fig. 3 shows original images from the "Image Parsing" set and automatically generated novel images with animated and reshaped humans and different types of backgrounds.

[1]CMU MoCap Database http://mocap.cs.cmu.edu/



(a) (b)

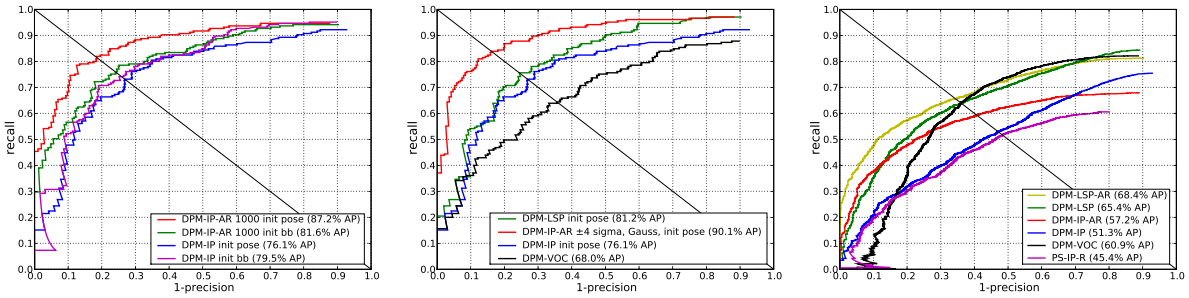
Figure 3: Examples of automatically generated novel images: (a) original image and (b) animated and reshaped synthetic samples with different backgrounds. Note the realism of the generated samples.

3. Articulated people detection

This section evaluates our data generation method for articulated people detection. For this we use the deformable part model (DPM) [11] and evaluate its performance on the "Image Parsing" dataset [25]. For training we use training sets from the publicly available datasets: PASCAL VOC 2009 (VOC) [10] consists of $2,819$ images of people captured over a wide range of imaging conditions; "Image Parsing" (IP) [25] consists of 100 images of fully visible people in a diverse set of activities such as sports, dancing, and acrobatics; the recently proposed "Leeds Sports Poses" (LSP) dataset [19] that includes $1,000$ images of people involved in various sports. We denote the models trained on these sets as DPM-VOC, DPM-IP and DPM-LSP. We introduce two new training sets obtained from IP by reshaping (R) and the combination of animating and reshaping (AR) training examples[2]. The models trained on this data *together* with the IP data are denoted DPM-IP-R and DPM-IP-AR accordingly. Average precision (AP) is used to compare performance and the PASCAL criterion [10] is used for matching.

DPM training. Training of DPM proceeds as usual [11]. However, we found that the initialization of DPM components significantly influences detection performance. I.e. the standard way to initialize the components based on the bounding box (BB) aspect ratio does not appear to be well suited for our task, as people with different poses often have similar BBs. We explore an alternative initialization strategy, where we cluster the images according to the relative displacement of the 2D joint locations w.r.t. the fixed body joint (neck joint in our case). The comparison of detection performance is presented in Fig. 4(a). DPM-IP-AR outperforms DPM-IP (81.6% vs. 79.5% AP) even when initialized by BB aspect ratios. Initializing DPM by pose clustering leads to an unequal distribution of training samples among different components and thus some components suffer from the lack of training data. This explains the

[2]The data is available for research purposes on our web page.



(a) (b) (c)

Figure 4: Comparison of different initializations for DPM components (a). Comparison of detection results of the DPM model on (b) "Image Parsing" and (c) multiscale "Leeds Sports Poses" datasets.

real/synthetic	AP, [%]
100 IP/0	76.1
100 IP/400 R	83.9
100 IP/400 AR	87.2
100 IP/900 AR	**88.6**
100 IP/1900 AR	88.1

Table 1: Results using "Image Parsing" (IP) data alone and jointly with Reshape (R) or Animate-Reshape (AR).

range	sampling	
	uniform	Gauss
$\pm 4\sigma$	85.4	**90.1**
$\pm 3\sigma$	**88.6**	85.1
$\pm 2\sigma$	88.0	83.2
$\pm 1\sigma$	85.6	87.3

Table 2: Results for different samplings of shape parameters in Animate-Reshape data.

performance decrease for DPM-IP (76.1% AP). However pose clustering accounts for a significant improvement for DPM-IP-AR (87.2% AP), as each component gets enough training data. This underlines the argument that our data generation method does indeed help to cover more shape and pose variations compared to the real data alone.

Data ratio. We study the influence of increasing shape and pose variations in the training data by changing the ratio between AR and IP data (results in Tab. 1). Clearly, performance is worst when training on IP data alone (76.1% AP). Adding 400 of R samples (increasing only shape variations) noticeably improves performance (83.9% AP). However adding the same number of AR samples (increasing both shape and pose variations) accounts for further improvements (87.2% AP). This supports the intuition that a global articulated people detector requires training data with large shape and pose variations and thus can significantly profit from our data generation method. Increasing the amount of AR data further improves the performance to 88.6% AP. Adding even more AR samples leads to a slight decrease in performance due to overfitting.

Shape variations. The ability to sample from the underlying 3D human shape distribution provides a direct control over generated data variability. Thus it is important to evaluate various ranges of shape changes and different sampling strategies. We sample shape parameters within ± 1, 2, 3 and 4σ (standard deviation) from the mean shape using uniform and Gauss-sampling and report the results in Tab. 2 for 100 IP/900 AR data. For both uniform and Gauss strategies sampling from $\pm 3\sigma$ outperforms $\pm 2\sigma$ as it better covers the space of possible shapes. Interestingly, by Gauss-sampling from $\pm 4\sigma$ and thus oversampling the tails of possible shape variations represented by our 3D human shape model we are able to improve the performance to 90.1% AP. Intuitively, the tails of the data distribution are important for learning powerful detectors. Increasing the sampling range increases the likelihood to sample unlikely but possible shape variations, which is far more difficult to achieve when using manually collected datasets only.

Summary of detection results. In Fig. 4(b) we summarize our findings and compare the obtained results to both DPM-VOC and DPM-LSP. DPM-VOC performs the worst (68.0% AP) trained on mostly upright people without strong articulations. This intuition is also supported by a better performance of DPM-IP (76.1% AP) trained from a much smaller set of images containing highly articulated people. Although training from a larger number of real samples (DPM-LSP) increases the detection rate (81.2% AP) this improvement is less pronounced compared to DPM-IP-AR (90.1% AP). This is due to the fact that the data variability is uncontrolled in LSP, as thus by adding more real samples we do not necessarily increase the variability. Training on our data generated from only 100 real images and having controllable pose and shape variations outperforms other models by a large margin achieving a remarkable 90.1% AP. We also show example detections at the equal error rate for DPM-IP-AR and DPM-IP in Fig. 5. Both qualitative and quantitative results clearly show the advantage of our method to increase the shape and pose variability of training data by sampling from the underlying 3D human shape distribution and changing human poses.

4. Articulated pose estimation

Motivated by the success of our data generation method to enable articulated people detection, this section proposes a new joint model for body pose estimation combining our

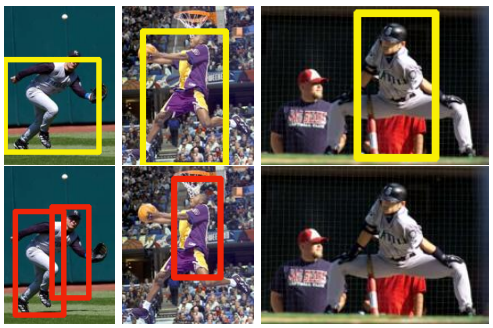


Figure 5: Examples of articulated people detections at EER by DPM trained on our joint synthetic and real "Image Parsing" (IP) data (top) and IP data alone (bottom). DPM trained on VOC2009 failed to detect people in these images.

pictorial structures model with DPM. We first briefly describe the Pictorial Structures (PS) model [14, 12] and then introduce our novel Joint PS-DPM model. We evaluate both models on the challenging "Image Parsing" dataset and show that pose estimation can directly profit from our strong articulated people detector. We use the percentage of correct parts (PCP) [13] measure for performance comparison.

4.1. Pictorial structures model

Pictorial structures (PS) [12, 14] represent the human body as a flexible configuration $L = \{l_0, l_1, ..., l_N\}$ of body parts. The state of each part i is denoted by $l_i = (x_i, y_i, \theta_i, s_i)$, where (x_i, y_i) gives the part position in image coordinates, θ_i the absolute part orientation, and s_i indicates the part scale relative to the part size in the scale normalized training set. Given image evidence E, the posterior of the part configuration L is described by $p(L|E) \propto p(E|L)p(L)$, where $p(L)$ is the kinematic tree prior and $p(E|L)$ is the likelihood of image evidence E for the body part configuration L. The tree prior describes dependencies between model parts and can be factorized as $p(L) = p(l_0) \prod_{(i,j)\in G} p(l_i|l_j)$, where G is the set of all directed edges in the kinematic tree, l_0 is assigned to the root node (torso) and $p(l_i|l_j)$ are pairwise terms along the kinematic chains. Pairwise terms are modeled by Gaussians in the transformed space of part joints while $p(l_0)$ is assumed to be uniform. The likelihood term is decomposed into the product of single part likelihoods $p(E|L) = \prod_{i=0}^{N} p(E|l_i)$,

We use our publicly available implementation [2]. In this implementation part likelihoods are modeled with AdaBoost classifiers [15] and image evidence is represented by a grid of shape context descriptors [5]. Inference is performed by sum-product belief propagation, which allows to compute marginal posteriors of each body part.

4.2. Joint PS-DPM model

While being conceptually similar, the DPM model and the PS model are designed for different tasks. The DPM model is designed for object detection and its parts are optimized to localize a bounding box of the person only. In particular it is non-trivial to map these parts to the locations of the anatomical body parts as is necessary for human pose estimation. On the contrary the PS model is defined directly in terms of anatomical parts and explicitly models their mutual positions and orientations. However anatomical body parts are not necessary optimal for detection, as they might be non-discriminative with respect to background.

To benefit from the complementary properties of PS and DPM models we define a joint model by embedding the evidence provided by DPM model into the PS framework. In the joint model we define the likelihood of the torso part as a product of two likelihood terms $p(E|l_i) = p_{ps}(E|l_i)p_{dpm}(E|l_i)$, where the first term is the original PS torso likelihood, and the second term is given by the torso prediction from the DPM. We adapt the DPM model to estimate the torso location by training linear regression model that predicts torso endpoints from the positions of the DPM model parts. These estimates are robust since the torso is typically associated with multiple parts of the DPM, which reduces uncertainty in the prediction. For each predicted torso location l_i we define $p_{dpm}(E|l_i) = \sigma(\mathrm{m}(l_i))$, where $\mathrm{m}(l_i)$ is the confidence score of the DPM detection, and $\sigma(\cdot)$ is a sigmoid function that calibrates the DPM score with respect to the PS likelihood. For all locations that did not have torso predictions we set the likelihood to $\varepsilon = 10^{-3}$.

4.3. Experimental evaluation

Here we evaluate both original PS and the proposed joint PS-DPM model on the task of pose estimation. In the following experiments the spatial and the part likelihoods of both models are learned on different training data, namely real "Image Parsing" (IP) data alone and together with the Reshape (R) data produced by our data generation method.

Training on IP data alone. First we report the best results obtained by training the PS model on IP data only. Similar to [2] we train part detectors on the training set augmented with the slightly rotated, translated and scaled versions of the original training samples. As in [2], we use a repulsive factor for lower and upper legs and perform inference by loopy belief propagation on the reduced state space of samples from part posteriors. Using IP data only we achieve 59.6% PCP. The results are shown in Tab. 3.

Training on IP and Reshape data. Our findings indicate that by jointly training on IP and Reshape data we improve over IP data alone. The best result is achieved by adding 1200 synthetic samples to the training data (61.9% PCP). Further increasing the proportion of Reshape samples leads to worse performance due to overfitting to synthetic samples, while decreasing the number of Reshape samples reduces variability and leads to worse performance. Training

Setting	Torso	Upper legs	Lower legs	Upper arms	Forearms	Head	Total
Image Parsing (IP)	84.9	71.5	61.5	50.2	36.6	71.2	59.6
+ Reshape (R)	87.8	75.1	65.9	52.4	36.1	71.7	61.9
+ Joint PS+DPM	**88.8**	**77.3**	**67.1**	53.7	36.1	73.7	63.1
Andriluka et al., [2] *	83.9	70.5	63.4	50.5	35.1	70.7	59.4
Yang&Ramanan, [31] *	82.9	69.0	63.9	55.1	35.4	**77.6**	60.7
Johnson&Everingham, [20]	87.6	74.7	**67.1**	**67.3**	**45.8**	76.8	**67.4**

* evaluated using our implementation of PCP criteria introduced in [13]

Table 3: Pose estimation results (PCP) on the "Image Parsing" (IP) dataset.

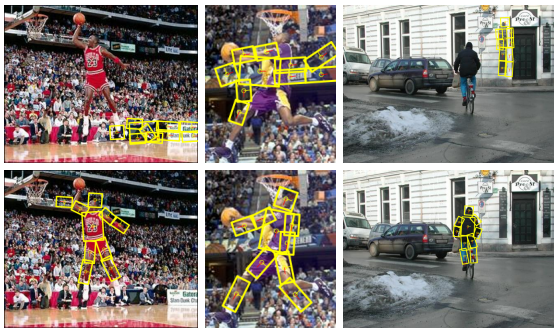


Figure 6: Comparison of body pose estimation results between the PS trained on IP (top) and our Joint PS+DPM model trained on IP + Reshape data (bottom).

on IP and Animate-Reshape (AR) data (60.0% PCP) performs slightly worse than the Reshape data. The PS does not benefit from the animated training data as it can already model such transformations via a flexible pose prior. The best performance is achieved by uniformly sampling from $\pm 1\sigma$ (61.9% PCP) while Gauss-sampling performs slightly worse ($\pm 2\sigma$, 61.2% PCP).

Training Joint PS-DPM. Results of training our Joint PS-DPM model on IP and AR data are shown in Tab. 3 (row 3). The Joint PS-DPM model outperforms the PS model alone (63.1% vs. 61.9% PCP). Expectedly, the localization of torso improved (87.8% vs. 88.8% PCP) which is explained by the increased confidence of torso estimation in the Joint PS-DPM model. Clear improvement is achieved for all body parts apart from forearms, while the limbs directly connected to the torso profit at most.

Comparison to the state of the art. We compare our results to the best results from the literature in Tab. 3. We outperform our previous work [2] and more complex discriminatively trained mixtures of parts model [31]. The achieved performance is slightly below [20] who use far more training data and learn *multiple* PS models after clustering similar poses. We envision that their clustering scheme could be effective in our case as well, in particular since we could generate sufficient amounts of training data even for clusters with rare poses. We leave this extension to future work.

Note that the results of [31] presented in Tab. 3 differ from those found in the original publication. The difference is due to the use of evaluation toolkit provided with the "Buffy" dataset [13], which deviates from the PCP criteria introduced in [13] in several ways leading to higher PCP scores[3]. For the sake of comparison we re-evaluate our method using the publicly available toolkit [13]. The results are shown in Tab. 4. Clearly, both peculiarities of evaluation procedure employed by [31] contribute to significantly higher PCP results.

In Fig. 6 we show examples of pose estimation results by our joint PS+DPM model trained on Reshape data and PS model trained on IP data alone. Note that the PS fails due to background clutter (left and middle) and presence of human-like structures (right). The Joint model uses additional information from the DPM torso prediction and thus is more robust. Clearly, correct estimation of torso position is the key to correct estimation of the rest parts.

5. Articulated pose estimation "in the wild"

Most recent work on articulated pose estimation considers a simplified problem by assuming that there is a single person in the image and that an approximate scale and position of the person is known [25, 19, 27, 30]. The proposed approaches typically output a single estimate of body configuration per image and do not provide any confidence score that the pose estimate is indeed correct. This ignores two important issues which arise when applying these approaches on real images. First, many images contain multiple people and so in addition to estimating poses of people it is also necessary to decide how many people are present. Second, for each person it becomes necessary to search over a wide range of possible positions and scales, and it is not clear how well the proposed methods are able to deal with such increase in complexity. We argue that in order to prop-

[3] According to the definition of PCP from [13] the body part is considered correct if *both* of its endpoints are closer to their ground truth positions than a threshold. The code in "Buffy" toolkit requires that the *average* over endpoint distances is smaller than the threshold. Such loose matching allows a segment to be accepted even if it is far from the ground-truth, because small distance of one endpoint can compensate for a large distance of the other endpoint. Another difference is that the code accepts multiple pose hypotheses as input, and evaluates the PCP score *only* for the hypothesis matching the ground-truth upper body bounding box. This is the "best case" evaluation that relies on the ground-truth annotation. In contrast, the PCP criteria [13] assumes there is one hypothesis for each part per image.

Setting	Torso	Upper legs	Lower legs	Upper arms	Forearms	Head	Total
Our method, our evaluation	88.8	77.3	67.1	53.7	36.1	73.7	63.1
Our method, loose matching	92.7	84.1	74.4	62.2	44.1	81.0	70.3
Our method, evaluation of [31]	**98.9**	**90.1**	**79.6**	68.8	48.1	92.5	**76.5**
Yang&Ramanan, [31], our evaluation	82.9	69.0	63.9	55.1	35.4	77.6	60.7
Yang&Ramanan, [31], loose matching	88.8	78.5	71.7	70.7	41.7	81.5	69.6
Yang&Ramanan, [31], evaluation of [31]	97.6	83.9	75.1	**72.0**	**48.3**	**93.2**	74.9

Table 4: Pose estimation results (PCP) on the "Image Parsing" (IP) when using our evaluation and evaluation of [31].

erly asses the state-of-the-art in articulated people detection and pose estimation it is necessary to consider these problems jointly. To that end we define a new dataset and evaluation criteria, and use them to validate the results obtained in Sec. 3 and Sec. 4 in a more realistic setting.

Dataset and evaluation criteria. The "Leeds Sport Poses" (LSP) dataset [19] contains images of people rescaled to the same scale and cropped around the person bounding box. We define a new dataset based on the LSP by using the publicly available original non-cropped images. This dataset, in the following denoted as "multi-scale LSP", contains 1000 images depicting multiple people in different poses and at various scales. We extended the annotations on the new dataset to include ground truth body configurations and bounding boxes of all people taller than 150 pixels resulting in $2,551$ annotated people total. To jointly asses the performance of detection and articulated pose estimation we evaluate the pose estimation in terms of recall and precision curves (RPC) and use AP to compare the performance. The PASCAL criterion [10] is used for matching people detections to the ground truth. For part matching to the ground truth we employ the PCP measure (Sec. 4) and use the people detector score as a confidence score of the hypothesis of each part. In addition to already mentioned training data we animate and reshape original LSP [19] training images (LSP-AR) and use them to train a DPM.

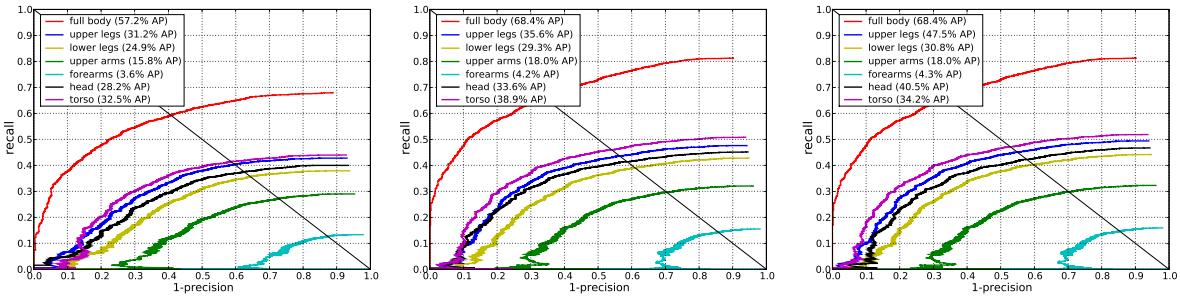
Results. Similar to [13] we use pre-filtering by running an articulated people detector. We collect all detections at the highest recall, and estimate poses independently for each of the detections matching the ground truth. All misdetections are considered when computing an RPC curve for each part.

We first evaluate the performance of DPM trained on different types of data. Results are shown in Fig. 4(c). Again DPM-IP-AR is much better than DPM-IP (57.2% vs. 51.3% AP), while DPM-LSP-AR outperforms DPM-LSP (68.4% vs. 65.4%) achieving the best result. These results show that the detectors trained on data augmented with reshaped and animated examples are more robust to strong pose variations. All DPM models outperform the PS model that is not trained discriminatively and is therefore more prone to failures in the presence of background clutter.

Fig. 7 shows RPC curves for individual body parts corresponding to different combinations of detection and pose estimation models. The best result is achieved by combining the DPM-LSP-AR detector with our Joint PS+DPM model (Fig. 7(c)). The performance varies greatly across parts. The localization is especially difficult for small parts such as forearms that are frequently occluded and foreshortened. To compare part detection performance across different models, we summarize the results in Tab. 5. Using DPM-VOC for pre-filtering achieves 18.2% AP, which is below PS-IP-R + PS (19.2% AP) performing better at high precision (cf. Fig 4(c)). DPM-IP-AR + PS achieves 21.2% AP. By using DPM-LSP-AR which is a better people detector we significantly improve the performance to 24.7% AP: localization of torso and head improves by more than 5% AP, while upper and lower legs improve by 4.4%. This clearly shows the importance of using a robust people detector to improve pose estimation of highly articulated people on multiple scales. Finally, DPM-LSP-AR + Joint PS+DPM achieves the best result (25.6% AP) outperforming other models for all parts. Torso, head and upper legs benefit most from better torso detection, as our joint model is able to detect the torso with higher confidence. The somewhat low overall results are due to a large number of partially occluded and strongly articulated people seen from untypical viewpoints. Our results indicate that even the currently best-performing methods often fail in such cases.

6. Conclusion

In this paper we propose a novel method for automatic generation of training examples from an arbitrary set of images. By using a 3D human shape model we generate realistic shape deformations of peoples' appearance. In addition, we animate reshaped samples by using a large set of motion capture data to generate plausible pose variations. We evaluate our data generation method for articulated people detection and pose estimation and show that for both tasks we significantly improve the performance when augmenting existing training data with our automatically generated images. In particular, we achieve very good results on the challenging "Image Parsing" benchmark using just 100 real images and a basic pictorial structures model. We also propose a joint model which integrates the evidence provided by DPM into the pictorial structures framework and experimentally show that the new model allows to further increase



(a) DPM-IP-AR + PS (b) DPM-LSP-AR + PS (c) DPM-LSP-AR + Joint PS+DPM

Figure 7: Detection results of different body parts on the multi-scale "Leeds Sport Poses" dataset.

Method	Torso	Upper legs	Lower legs	Upper arms	Forearms	Head	Total
DPM-LSP-AR + Joint PS+DPM	**40.5**	**37.5**	**30.8**	**18.0**	**4.3**	**34.2**	**25.6**
DPM-LSP-AR	38.9	35.6	29.3	18.0	4.2	33.6	24.7
DPM-IP-AR + PS	32.5	31.2	24.9	15.8	3.6	28.2	21.2
DPM-VOC + PS	29.9	25.2	20.0	14.2	3.6	27.4	18.3
PS-IP-R + PS	29.1	28.7	23.5	14.7	4.0	24.5	19.5

Table 5: Average precision (AP) of part estimations by different methods on multi-scale "Leeds Sport Poses" dataset.

the performance. Finally we propose a new challenge of joint detection and pose estimation of multiple articulated people in cluttered sport scenes.

Acknowledgements. We would like to thank Alec Jacobson and Pravin Bhat for helpful discussions.

References

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

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

[3] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis. SCAPE: Shape completion and animation of people. In *ACM TOG (Proc. SIGGRAPH'05)*.

[4] C. Barnes, E. Shechtman, A. Finkelstein, and D. B. Goldman. PatchMatch: A randomized correspondence algorithm for structural image editing. *SIGGRAPH'09*.

[5] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *PAMI'02*.

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

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

[8] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: A benchmark. In *CVPR'09*.

[9] M. Enzweiler and D. M. Gavrila. A mixed generative-discriminative framework for pedestrian classification. In *CVPR'08*.

[10] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL visual object classes (VOC) challenge. *IJCV'10*.

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

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

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

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

[15] Y. Freund and R. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *JCSS'97*.

[16] N. Hasler, C. Stoll, M. Sunkel, B. Rosenhahn, and H.-P. Seidel. A statistical model of human pose and body shape. In *CGF (Proc. Eurographics'08)*.

[17] C. Ionescu, L. Bo, and C. Sminchisescu. Structural SVM for visual localization and continuous state estimation. In *ICCV'09*.

[18] A. Jacobson, I. Baran, J. Popović, and O. Sorkine. Bounded biharmonic weights for real-time deformation. In *SIGGRAPH'11*.

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

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

[21] J. Marin, D. Vazquez, D. Geronimo, and A. Lopez. Learning appearance in virtual scenarios for pedestrian detection. In *CVPR'10*.

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

[23] R. Okada and S. Soatto. Relevant feature selection for human pose estimation and localization in cluttered images. In *ECCV'08*.

[24] L. Pishchulin, A. Jain, C. Wojek, M. Andriluka, T. Thormählen, and B. Schiele. Learning people detection models from few training samples. In *CVPR,'11*.

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

[26] C. Rother, V. Kolmogorov, and A. Blake. "grabcut": interactive foreground extraction using iterated graph cuts. *ACM Trans. Graph.'04*.

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

[28] G. Shakhnarovich, P. Viola, and T. Darrell. Fast pose estimation with parameter sensitive hashing. In *ICCV'03*.

[29] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from a single depth image. In *CVPR'11*.

[30] Y. Wang, D. Tran, and Z. Liao. Learning hierarchical poselets for human parsing. In *CVPR 2011*.

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

[32] S. Zhou, H. Fu, L. Liu, D. Cohen-Or, and X. Han. Parametric reshaping of human bodies in images. *SIGGRAPH'10*.