Qualitative Pose Estimation by Discriminative Deformable Part Models

Hyungtae Lee, Vlad I. Morariu, and Larry S. Davis

University of Maryland, College Park

Abstract. We present a discriminative deformable part model for the recovery of qualitative pose, inferring coarse pose labels (*e.g.*, *left, front-right, back*), a task which we expect to be more robust to common confounding factors that hinder the inference of exact 2D or 3D joint locations. Our approach automatically selects parts that are predictive of qualitative pose and trains their appearance and deformation costs to best discriminate between qualitative poses. Unlike previous approaches, our parts are both selected and trained to improve qualitative pose discrimination and are shared by all the qualitative pose models. This leads to both increased accuracy and higher efficiency, since fewer parts models are evaluated for each image. In comparisons with two state-of-the-art approaches on a public dataset, our model shows superior performance.

1 Introduction

The analysis of human actions in videos is an important task for many computer vision applications. In general, actions can be inferred from the kinematic movements of a person's limbs. However, people are highly articulated, limbs occlude each other, loose clothing conceals shape, and low resolution or motion blur lose informative features such as edges. All of these conditions confound pose estimation, making it a difficult and widely studied problem. To ameliorate these problems, researchers often introduce additional information, *e.g.*, multi-view images or depth information, if available, to reconstruct pose in 3D coordinates [1–8]. Instead of relying on additional information, our approach is based on the observation that for many action recognition tasks, it may be sufficient to infer only a qualitative description of a pose, *e.g.*, 'bent', 'laying down', 'stretching', 'crouched', 'facing left', 'facing right', even if the exact joint locations are not identified. We expect models that infer a qualitative description of a person's pose to be more robust in the presence of the problems that confound exact joint localization. We propose a model for inferring qualitative pose labels automatically from single monocular images.

We model each qualitative pose by a root filter and set of deformable parts, similar to [9]. Unlike the problem of human detection, for which part deformation is modeled to introduce invariance to slight pose changes, the problem of qualitative pose estimation benefits from models that exploit part appearance and deformation to discriminate between pose changes. While we do not require

K.M. Lee et al. (Eds.): ACCV 2012, Part II, LNCS 7725, pp. 734–747, 2013.

Fig. 1. Part appearance provides information that can be used to discriminate between qualitative poses. In the first row, the figure shows various qualitative poses. The second and third rows show patches cropped to contain two sets of target joints, the combination of head-chest-right elbow (row 2) and left elbow-hip-left knee (row 3). We note that appearance of patches covering the same joints varies according to qualitative poses. Our approach takes advantage of this relationship between part patch appearance and qualitative pose.

exact recovery of joint locations, it is important for part models to provide information that can be used to discriminate between qualitative poses. For this reason, when training part models, we ensure that models are tightly clustered in pose space (similar to Poselets [10]), and train multiple models covering the same physical parts in various configurations and viewing angles (see Figure 1), allowing the relevance of each part model to be automatically adjusted for each qualitative pose to improve discrimination. Given trained root and part filters, we optimize appearance and deformation weights for each qualitative pose and train a multi-class model that fuses the outputs of each qualitative pose model.

Our main contribution is a qualitative pose detection approach based on state-of-the-art deformable part models, with parts that are automatically initialized and trained on semantically meaningful pose clusters that are more discriminative than those initialized by random [10] or greedy [9] selection (in the latter, parts are selected to maximize the energy of the corresponding root filter subregion). A nice property of our approach is that the same parts are shared by all qualitative pose models (but are incorporated into each model using different weights), which requires that the computationally expensive sliding

window search to be performed only once for each part. We demonstrate the performance of our approach on a public database and compare against two baseline approaches from [9] and [11].

In section 2, we discuss related work. In section 3, we detail our proposed model. In section 4, we present the experimental results that demonstrate the performance of our approach. We present our concluding remarks in section 5.

2 Related Work

The literature on pose estimation is vast and includes methods that extract 2D pose using part models [12–20] and that estimate 3D pose from single or multiple views [1–8]. We focus our discussion on 2D pictorial structure methods as they are most related to our work. Pictorial structure models were introduced by Fischler and Elschlager [12] to represent objects as a collection of parts connected by spring-like connections. Parts encode local appearance and their locations can vary subject to a specified deformation cost. While a straightforward search for the optimal locations of parts is computationally expensive, the search becomes practical under certain conditions. For example, Felzenszwalb and Huttenlocher [13] showed that if the pictorial structure is acyclic, and the relationships between pairs is expressed in a restricted form, the generalized distance transform can be used to compute the globally optimal configuration at a cost that is linear in the number of part locations. In subsequent work, Felzenszwalb et al. [9] proposed a more general deformable part model, consisting of roots, parts, and deformation costs which are all discriminitively trained using Latent SVM. Part locations are automatically initialized from an initial estimate of the root filter by a greedy cover of high energy areas of the root filter, and their deformations are optimized efficiently using the generalized distance transform.

Bourdev et al. introduce a novel concept of parts, Poselets, which are tightly clustered in both appearance and configuration space [21, 10]. As proposed, these parts do not necessarily coincide with body segments (*e.g.* upper arm, lower arm, torso), but generally capture combinations of portions of parts which are distinctive in certain views (*e.g.*, frontal face and right-shoulder combination). During training, candidate Poselets are obtained by repeatedly selecting a random patch in a training image, finding patches in other images which are near in configuration space, and training a Poselet detector using these patches. At test time, Poselet detections (or *activations*) are obtained by multi-scale sliding window search, and objects are detected either by Max Margin Hough Voting [21] or by clustering mutually consistent activations [10]. An attractive feature of Poselets is that they can easily propagate additional labels that were provided with the initial training set, *e.g.*, segmentation masks and joint locations. Consequently, Poselets have been applied to various problems, including the estimation of segmentations [22], actions [23, 11], subordinate categorization [24], and attribute classification [25]. Poselets have also been incorporated into pictorial structure models for 2D pose estimation [26], with Poselets organized into a hierarchy of various sizes, covering individual parts, combinations of parts, and even the entire body.

Several exististing approaches predict discretized viewing directions (which fits our definition of qualitative pose) as part of their frameworks. Andriluka et al. [4] train eight independent view-point specific pictorial structure based detectors, whose outputs are combined using a linear SVM. Each model is trained independently and uses standard body parts (head, torso, upper/lower legs, upper/lower arms, feet). Maji et al. [11] predict discretized viewing directions as part of a static action recognition framework using the Poselet framework. To achieve this, they train 1200 Poselets, which are applied at test time.

Our approach builds on deformable part models [9] and Poselets [10], but they are optimized for the purpose of qualitative pose estimation instead of object detection. Instead of selecting Poselets by random selection [10] or greedy cover [9], our approach automatically selects clusters from sets of joints whose variations are predictive of qualitative pose. We train multiple models, one for each qualitative pose, allowing each model to select part deformation weights that best allow for discrimination between qualitative poses. Our approach is trained to maximize discrimination at all levels (parts, deformation weights, combination of output), unlike [4], and requires few part models (in our experiments we used only 64 parts, while [11] employed 1200).

3 Qualitative Pose Estimation

The block diagram of our approach is shown in Figure 2. To reduce overfitting, we divide the training dataset into two sets; one is for training root and part filters and the other is for training Q(Qualitative)-Pose models by Latent SVM [9] and calibrating them to each other. Root regions, which are defined as fixed aspect ratio bounding boxes whose vertical extents are defined by the head and the waist of a person, are first cropped from training images and are divided into sets according to their labeled qualitative pose. Root filters are learned via SVM with HOG features constructed from the collected images in each set. Part filters cover a combination of joints, which are selected manually based on how predictive their appearance variations are of qualitative pose. In our experiments, parts are defined by three joints: head - upper torso - right elbow, head - upper torso - left elbow, right elbow - lower torso - right knee, and left elbow - lower torso - left knee. For each part, training images are divided into clusters by k-means clustering according to the similarity of joint configurations. Next, part filters are learned as for the root filter. The root and part filters are then applied to the second (held out) set of training images, and the set of activations are used to train the weights of a Latent SVM model that detects qualitative poses based on root and part filter activations. During testing, we first extract activations of trained root and part filters by sliding window search. Then, for each qualitative pose model, we select an activation for each part to maximize the joint model score, and then apply the linear model learned by multi-class SVM to predict the best matching qualitative pose.

We describe the training process in more detail in the next subsections. In section 3.1, we provide the model formulation. We then describe root and part filter training and model parameter optimization in section 3.2 and 3.3, respectively.

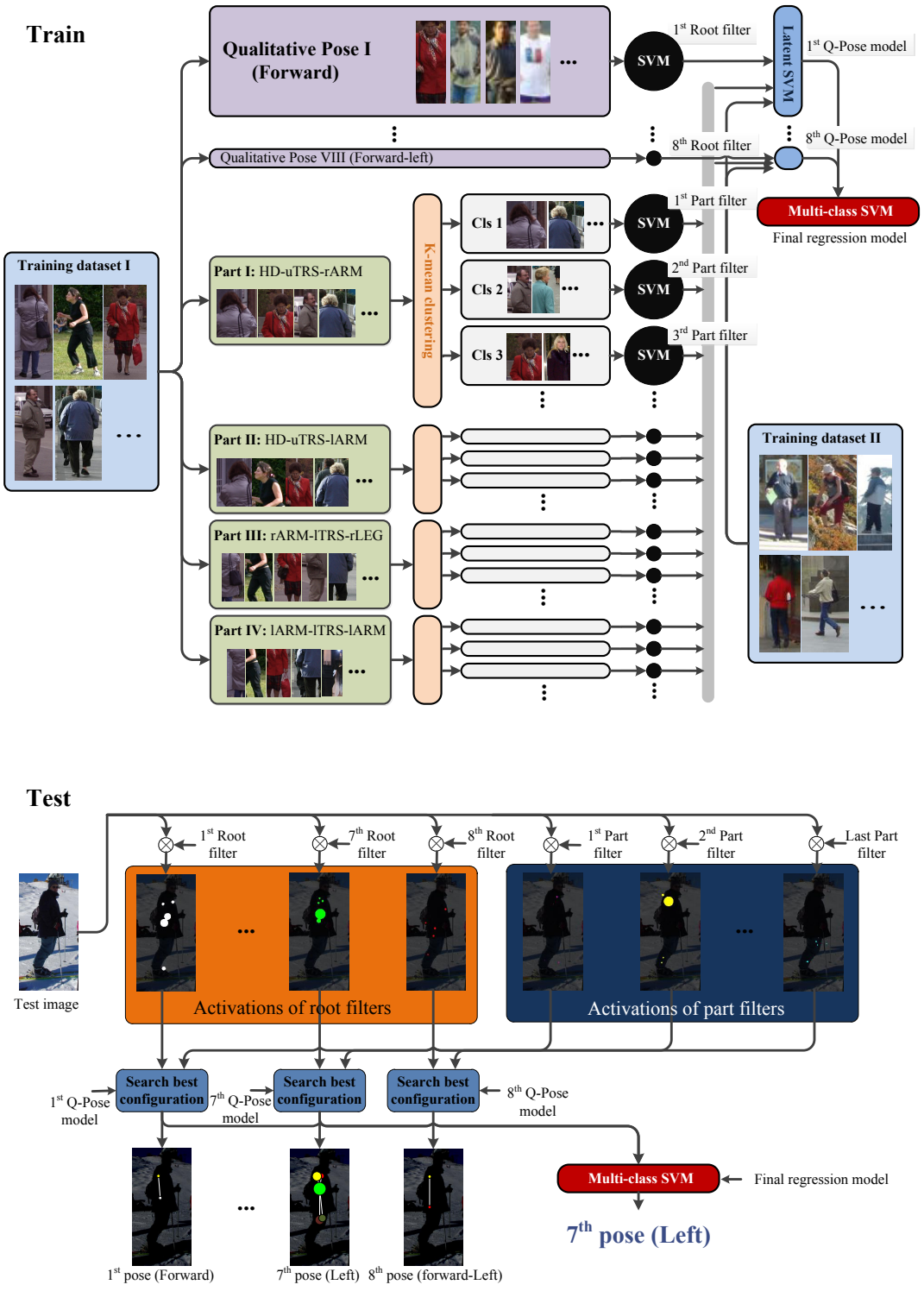

Fig. 2. Overview of training and test procedure

3.1 Model Formulation

Let $q_i,\ i = 1, 2, \cdots, Q$ denote the set of qualitative poses. A model for each qualitative pose, $M_i = \{r_i,\ P,\ A_i,\ w^i_{0:K},\ \boldsymbol{w}^i_{d;1:K}\}$ is trained, where r_i is the root filter for qualitative pose i, $P = \{p_1,\ p_2,\ \cdots,\ p_K\}$ is a set of part filters,

$A_i = \{a_1^i,\ a_2^i,\ \cdots,\ a_K^i\}$ is a set of anchor positions which specify the relative position of k^{th} part to the root, and $w_{0:K}^i$ and $\boldsymbol{w}_{d;1:K}^i$ are model parameters that weigh appearance and deformation costs, respectively ($\boldsymbol{w}_d$ is a vector that defines the deformation cost as in [9]). Every model uses the same set of part filters, P. Q and K denote the size of the set of qualitative poses and parts, respectively. At test time, filter activations generate a set of candidate locations for each part. A hypothesized qualitative pose is formed by selecting one of the candidate part locations for each of the root and each of the K parts, $L_i = \{l_0^i,\ l_1^i,\ \ldots,\ l_K^i\}$. We model the probability $p\,(L_i|I, M_i)$ that the configuration is L_i given the image I and the model M_i and decompose it as follows:

$$p(L_i|I, M_i) \propto p(I|L_i, M_i)p(L_i|M_i). \tag{1}$$

The distribution $p(I|L_i, M_i)$ measures the likelihood of fitting the model to a particular image given a part configuration, and $p(L_i|M_i)$ is the prior distribution that each part is placed at a particular location. The best configuration L_i can be obtained by MAP estimation as

$$L_i^* = \underset{L_i \in L_a(I)}{\arg\max}\ p(L_i|I, M_i). \tag{2}$$

The likelihood of configuration L_i is modeled as the product of the i^{th} root likelihood and the individual part likelihoods,

$$p(I|L_i, M_i) = p(I|r_i, l_0^i, w_0^i) \prod_{k=1}^{K} p(I|p_k, l_k^i, w_k^i). \tag{3}$$

where $p(I|p_k, l_k^i, w_k^i) = exp(-w_k^i m_k(I, l_k^i))$, and $m_k(I, l)$ measures the response of the k^{th} filter at position l in image I.

The prior distribution of the configurations can be expressed as a product of a root location prior and part location priors given the root location,

$$p(L_i|M_i) = p(l_0^i|M_i) \prod_{k=1}^{K} p(l_k^i|l_0^i, a_k^i, \boldsymbol{w}_{d;k}^i). \tag{4}$$

The prior distribution of root location, $p(l_0^i|M_i)$ is modeled as a uniform distribution, and $p(l_k^i|l_0^i, a_k^i, \boldsymbol{w}_{d;k}^i) = exp(-\boldsymbol{w}_{d;k}^{i^T} f_d(l_k^i, l_0^i + a_k^i))$ is the probability that the k^{th} part is placed at l_k^i given the root location. The deformation function $f_d(l_i, l_j) = \left[-dl_x\ -dl_y\ -dl_x^2\ -dl_y^2\right]^T$, where $dl = l_i - l_j$, is defined as in [9].

The score of a hypothesis is defined as the negative logarithm of equation 1,

$$score(I, L_i) = w_0^i m_0^i(I, l_0^i) + \sum_{k=1}^{K} w_k^i m_k(I, l_k^i) + \sum_{k=1}^{K} \boldsymbol{w}_{d;k}^{i^T} f_d(l_k^i, l_0^i + a_k^i). \tag{5}$$

which can be more compactly represented as the dot product, $W_i^T \Phi(I, L_i)$, of model parameters W_i and a vector $\Phi(I, L_i)$ specifying a matching score and deformation cost of each part in its own location,

$$W_i = \left[w_0^i; \ldots; w_K^i; \boldsymbol{w}_{d;1}^i; \ldots; \boldsymbol{w}_{d;K}^i \right],$$
$$\Phi(I, L_i) = \left[m_0^i(I, l_0^i); \ldots; m_K(I, l_K^i); f_d(l_1^i, l_0^i + a_1^i); \ldots; f_d(l_K^i, l_0^i + a_K^i) \right]. \quad (6)$$

3.2 Training Root and Part Filters

We define a root that represents the general position of the entire human and provides an anchor position for each part (this anchor position will vary with qualitative pose). The root is defined by the head and the waist (the width of the box is a fixed ratio of the height, which is defined as the distance between the head and the waist), and its appearance and position does not greatly change with various human poses or actions. For each qualitative pose, a root filter is trained to model the general location of parts. The part anchor positions are computed by averaging relative positions of each part to the root in all training images labeled as the specified qualitative pose. To train a root model of each qualitative pose, we collect examples cropped around the root region from images labeled as a particular qualitative pose. We crop and resize each example to a fixed height and aspect ratio. The height is set to the median value of every cropped root region and the width is calculated by dividing the height by the fixed aspect ratio. Given these examples, we extract HOG features from the collected positive examples and randomly select ten times as many negatives and train linear SVM classifiers to discriminate between positive and negative examples. As for Poselet training, we scan over background images that contain no people, collect false positive examples, and retrain linear SVM classifiers, repeating this process a few times to train the classifier efficiently with a large number of negative examples. We note that each hypothesis has one root filter.

To ensure that part filters can be used to discriminate between qualitative poses, we select parts by clustering combinations of joints that are expected to vary in predictable ways with respect to qualitative pose. Figure 3 shows how parts composed of certain joint triples exhibit a large spatial and appearance variation with respect to qualitative pose. We define our parts by clustering the configurations of combinations of three joints, which in our case define pairs of limbs: head - upper torso - right elbow, head - upper torso - left elbow, right elbow - lower torso - right knee, and left elbow - lower torso - left knee.

During part filter training, images are first resized to have root regions of the same size. For each joint triplet, training examples are then selected by cropping a region containing the three selected joints. For each combination of joints, training samples are divided into n classes by k-means according to similarity of joint configuration. To cluster part training examples, the joint configuration of each part is represented as a vector of concatenated joint positions relative to the torso, and the similarity of joint configurations between two examples is computed by Euclidean distance between the configuration vectors. Each training sample is resized to the median size of the training example. As a result of this process, each joint triplet generates n parts which correspond to clusters in joint configuration space.

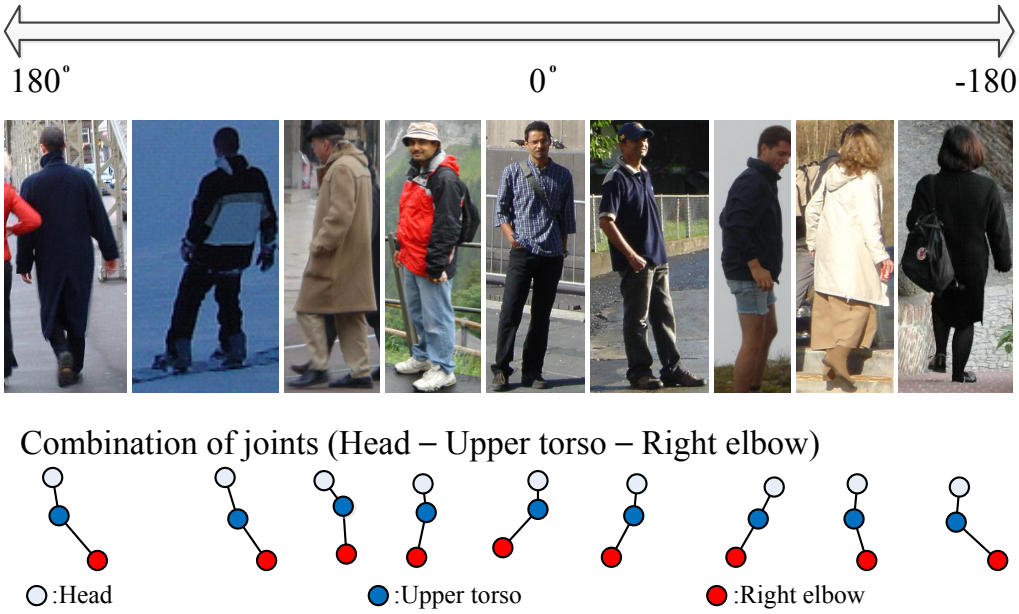

Fig. 3. The overall appearance and arrangement of the head, upper torso, and right elbow joints varies significantly with qualitative pose, as shown by the sample images and illustration of the three corresponding nodes. For this reason, the three joints together can be considered a discriminative part.

Note that a part, if detected, does not directly imply a certain qualitative pose, but we expect that some parts are more predictive of certain qualitative poses than others (our experimental results confirm this). Given the selected training samples for each part, we train part filters in a similar way to the root filter. The only difference is that for training the parts we use negative examples extracted from images in other clusters from the training set while for training the root we extract them from background images. By including samples from other part clusters as negative part samples, we train part filters that better discriminate between joint configurations. After root and part filter training, we obtain Q root filters and $4n$ part filters.

3.3 Learning Model Parameters

Each image in the training dataset is labeled with its qualitative pose. We indicate the training dataset as $\{I_n, b_n\}_{n=1}^{N}$, where I_n is an image and b_n is its label. Given an image and its label, the trained root and part filters can be applied to detect candidate locations of parts. For every qualitative pose, we learn a deformable part model over the root and part filters using the latent SVM formulation [9]. For each qualitative pose, a classifier that scores an image I is defined as

$$f_W(I) = \max_{L \in Z(I)} W^T \Phi(I, L), \tag{7}$$

where $Z(I)$ is the set of all possible combinations of activations (here, we only consider the locations corresponding to root and part filter activations). Model parameters can be learned by minimizing the objective function

$$L_D(W) = \frac{1}{2}||W||^2 + C\sum_{i=1}^{M}\max(0, 1 - y_i f_W(I_i)), \tag{8}$$
$$\text{where } y_i = \begin{cases} 1 & \text{if } I_i \text{ is a positive example} \\ -1 & \text{otherwise.} \end{cases}$$

The standard hinge loss, $\max(0, 1 - y_i f_W(I_i))$ is concave when an image I_i is labeled as positive because the classifier, $f_W(I)$ is convex. Latent SVM optimization specifies the latent value L^* for every positive image and yields a linear form,

$$f_W(I) = W_t^T \Phi(I,\ L^*), \tag{9}$$
$$\text{where } L^* = \arg\max_{L \in Z(I)} W_{t-1}^T \Phi(I, L).$$

While searching for the best configuration L^*, the algorithm uses the parameter W_{t-1} learned in the previous step. In other words, the semi-convex optimization is solved by repeatedly optimizing two separate convex functions, a process called "coordinate descent". The first part of the optimization involves computing the overall score of each configuration of root and parts and selecting the highest scoring configuration. In our case, $Z(I)$ is a small enough set for all candidate configurations to be considered in a reasonable amount of time. In the second part of the optimization, we compute the model parameter W_t using a linear SVM.

We consider all configurations in images labeled as other qualitative poses as negative examples. To avoid considering unlikely configurations as negative examples, we collect only the best configuration for each root activation. Because negative examples are very numerous compared with positive examples, we extract a set of hard negative examples in every iteration of optimization, and ignore the remaining negatives during that iteration.

4 Experiments

We evaluate our framework on the public INRIA pedestrian database [27], which consists of images that contain upright pedestrians with annotated bounding boxes. Our aim in these experiments is to recognize qualitative poses by analyzing the entire body, so we did not consider datasets such as PASCAL VOC database which contain many images in which people are often only partially visible. While the INRIA pedestrian database might be considered easy for the task of human detection, it is a difficult dataset for the task of determining qualitative pose (as our experiments will show). To evaluate our approach, we assume that the person has been roughly localized, using a detector such as that

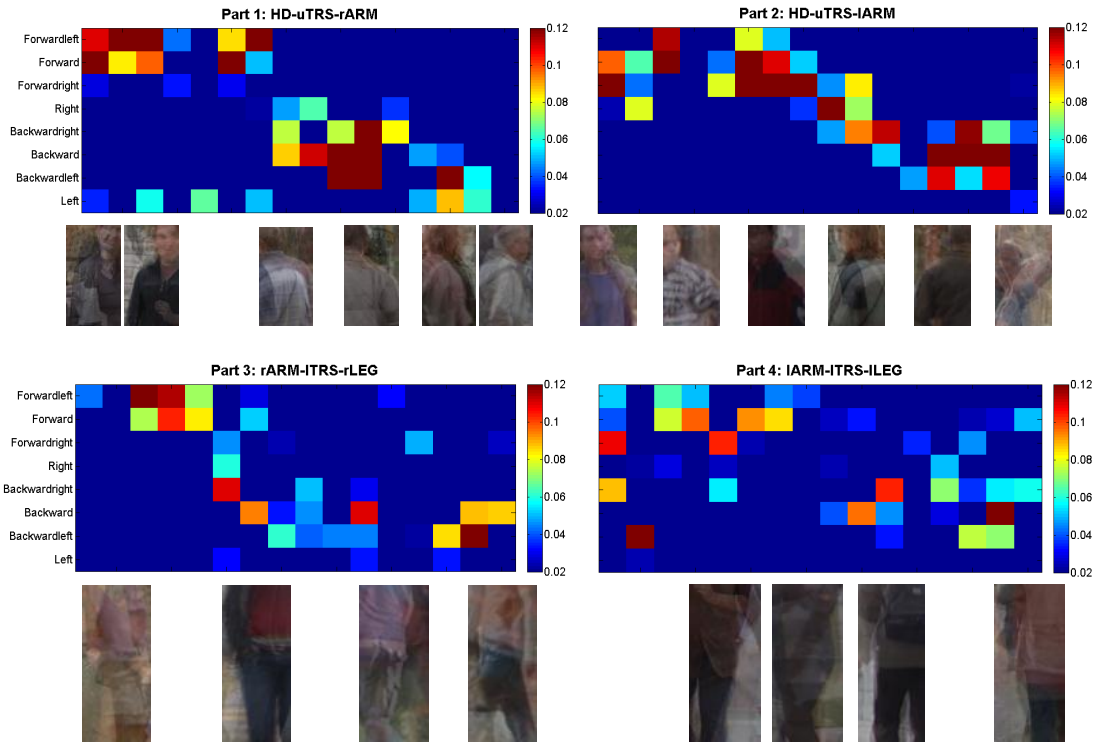

Fig. 4. The weights obtained by the optimization in equation 9 for each of the 8 qualitative pose models (y-axis) and each of the 16 part models (x-axis)

of Felzenszwalb et al. [9], so we focus only on assigning a qualitative pose label to regions extracted around the annotated bounding boxes. To increase the effective size of our training set, we also flip images along the vertical axis. Since bounding boxes may exclude part of a person region due to annotation errors, we cropped images only after adding a suitably large amount of padding to the human bounding boxes. We split the database randomly into three sets using a ratio of 2:2:1; the first split is for training part and root filters, the second is for validation, and the last is for testing. We discretize the qualitative pose into 8 discrete bins of angles corresponding to the direction that a person's torso is facing with respect to the camera, and so construct 8 qualitative pose models. Each bin covers 45 degrees.

To train root and part filters, we labeled the head, neck, waist, elbows, and knees, and specified the qualitative pose of each image. While training part filters, we set n, the number of clusters obtained from applying k-means of training samples to 16. We use 200 positive examples and 2000 negative examples for training each filter. We extract false positives and retrain for ten iterations.

Figure 4 shows the appearance weights obtained by the optimization in equation 9 for each of the 8 qualitative pose models. The qualitative poses are along the y-axis, ordered in circular fashion from forward-left to left. The part filters trained on the clustered joint triples are listed on the x-axis. These are also roughly ordered by the distribution of the qualitative pose labels of the training images belonging to each cluster, so that parts are also ordered circularly

Table 1. AUCs of each approaches. F, B, L, and R abbreviate 'forward', 'backward', 'left', and 'right', respectively. (The best result in each pose is in bold font.)

	BL	L	FL	F	FR	R	BR	B
Felzenswalb et al. [9]	0.6246	0.6682	0.6222	0.7868	0.5656	0.6526	0.6659	0.7414
Maji et al. [11]	0.5910	0.7186	0.6581	0.7786	0.7154	0.6147	0.6154	0.7423
Our approach	**0.7606**	**0.8542**	**0.7988**	**0.8963**	**0.7790**	**0.8090**	**0.8270**	**0.8967**

from forward-left to left. As expected, the strong diagonal weights in many of these images show that the parts obtained by our approach are indeed predictive of qualitative pose. Conversely, there still remains enough confusion that it is necessary to combine evidence from the multiple parts. We conclude that while upper parts (part 1 and 2) are more associated with qualitative poses, lower parts (part 3 and 4) include variations that are caused by other sources in addition to qualitative pose.

To evaluate our performance, we implement two state-of-the-art approaches for our qualitative pose estimation problem. As for our approach, pose-specific models are first trained independently on a training subset using the deformable part-based model (DPM) [9] and Poselets [10, 11] using the same training, validation, and test partitions. Independent model scores are calibrated against each other on a validation set by a multiclass SVM. We applied the DPM training/testing code as provided by the original authors, with a modified input training set (the 8 qualitative poses), a single component instead of mirrored left-right models (we care about facing direction), and a subsequent multiclass calibration step. We also compare to the Poselet code of Bourdev et al. [10], but since the training code is not provided by the authors, we implement the training procedure described in [10]. We use a pose activation vector that collects detection scores of 1200 Poselets as the pose representation, as in [11]. Figure 5 shows the performance of each independent qualitative pose model and compares our approach with the other two alternatives on INRIA pedestrian database. Based on the ROC curves, our approach outperforms the other methods for every qualitative pose. Table 1, which shows the area under the ROC curves (AUC), also shows that our approach outperforms the alternatives.

Figure 6 shows the confusion matrix of the three approaches, obtained after the independent model outputs are combined using a multi-class SVM. The confusion matrix for our approach has a much more pronounced diagonal than the other two alternatives, which is expected, given the individual qualitative pose detection performance. As one would expect, there is a lot of confusion between neighboring poses. Commonly, 'forward' and 'backward' are well detected but subtle differences between right- or left-facing poses are often misclassified. This has also been observed by other researchers [11], who have noted that human perceptual ability also distinguishes between cardinal directions (front, back, left, right) direction better than others such as front-right, backward-left, etc.

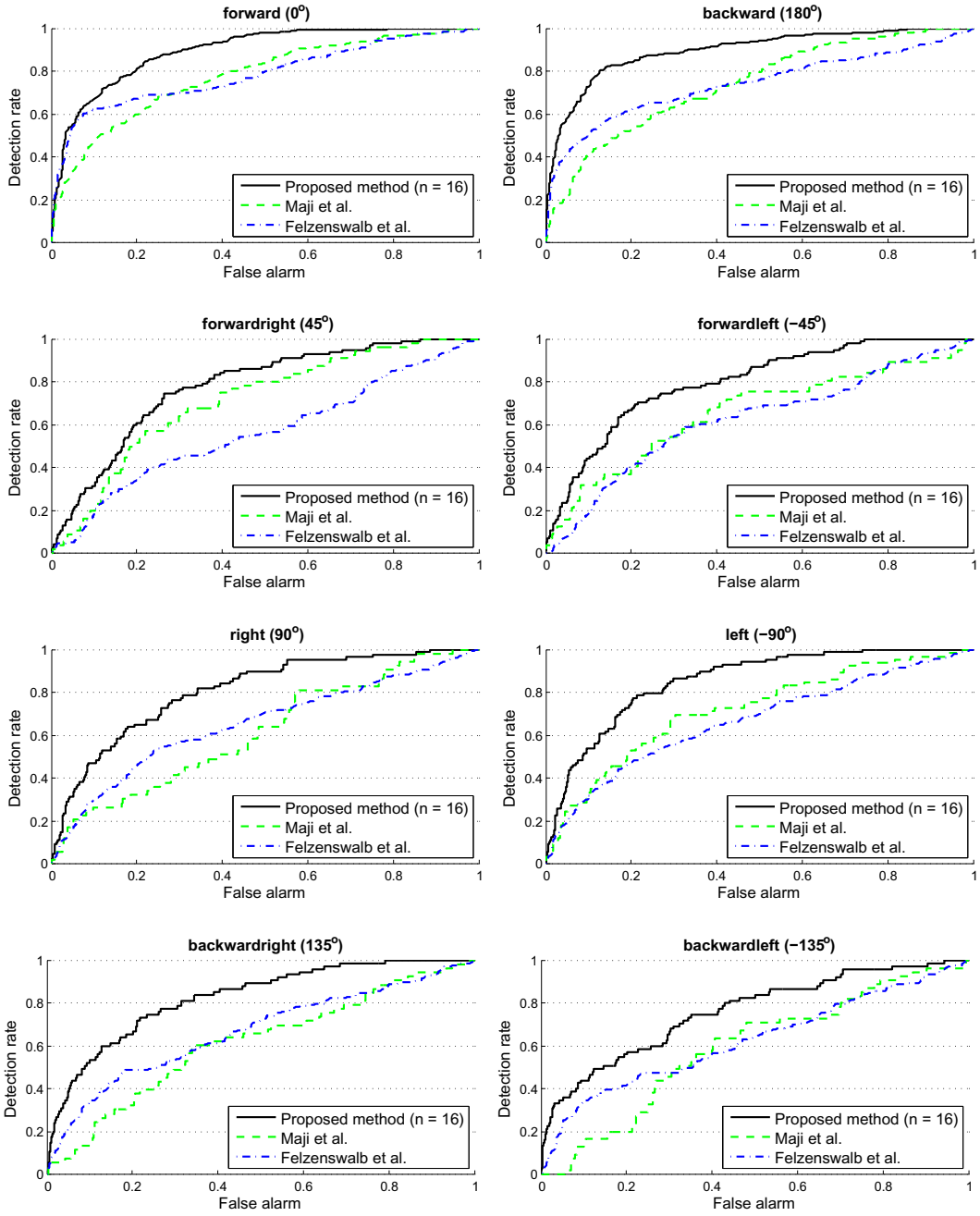

Fig. 5. ROC curves for performance of each qualitative pose model on the INRIA person database

While [9] and [11] achieve different performance between qualitative poses, our approach maintains a consistent level of detection for every class. Table 2 shows the overall recognition rate. Errors are computed by a mean squared error from misclassified class to groundtruth, where the distance between front and front-right is 1, front and right is 2, and so on. Our approach outperforms the others using these measures, as well.

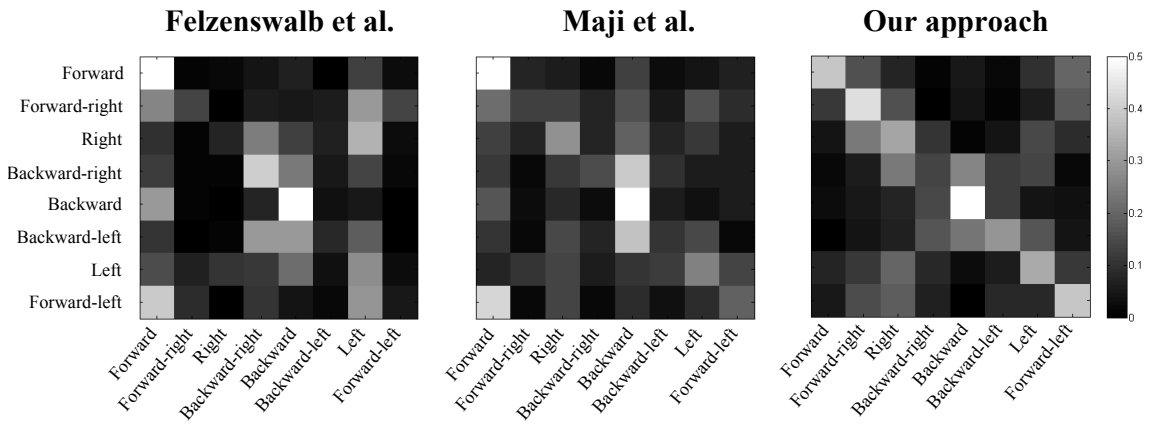

Fig. 6. Confusion matrix of three approaches. **Left:** Felzenswalb et al. [9], **Center:** Maji et al. [11] and **Right:** our approach.

Table 2. Overall recognition results of three approaches on the INRIA pedestrian database. (Bold font indicates the best result.)

	Recog. rate	Errors
Felzenswalb et al. [9]	0.2909	1.9868
Maji et al. [11]	0.2814	1.9431
Our approach	**0.3485**	**1.6810**

5 Conclusions

We presented a qualitative pose estimation approach that is based on discriminative deformable part models. Unlike previous approaches, we give special attention to the selection of part models, replacing random selection and greedy cover steps with an automatic clustering of part poses. The part appearance and deformation parameters are trained discriminatively for each qualitative pose model, and the outputs of all pose models are combined using a multi-class classifier. Our approach shows improved performance on the INRIA pedestrian database against two state-of-the-art approaches.

References

1. Hofmann, M., Gavrila, D.M.: Multi-view 3D human pose estimation combining single-frame recovery, temporal integration and model adaptation. In: CVPR (2009)
2. Wei, X.K., Chai, J.: Modeling 3D human poses from uncalibrated monocular images. In: ICCV (2009)
3. Guan, P., Weiss, A., Balan, A.O., Black, M.J.: Estimating human shape and pose from a single image. In: ICCV (2009)
4. Andriluka, M., Roth, S., Schiele, B.: Monocular 3D pose estimation and tracking by detection. In: CVPR (2010)
5. Daubney, B., Xie, X.: Tracking 3D human pose with large root node uncertainty. In: CVPR (2011)

6. Gall, J., Yao, A., Van Gool, L.: 2D Action Recognition Serves 3D Human Pose Estimation. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part III. LNCS, vol. 6313, pp. 425–438. Springer, Heidelberg (2010)
7. Lonescu, C., Li, F., Sminchisescu, C.: Latent structured models for human pose estimation. In: ICCV (2011)
8. Chen, C., Heili, A., Odobez, J.: Combined estimation of location and body pose in surveillance video. In: AVSS (2011)
9. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part based models. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) (2010)
10. Bourdev, L., Maji, S., Brox, T., Malik, J.: Detecting People Using Mutually Consistent Poselet Activations. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part VI. LNCS, vol. 6316, pp. 168–181. Springer, Heidelberg (2010)
11. Maji, S., Bourdev, L., Malik, J.: Action recognition from a distributed representation of pose and appearance. In: CVPR (2011)
12. Fischler, M., Elschlager, R.: The representation and matching of pictorial structures. IEEE Transactions on Computer 22 (1973)
13. Felzenszwalb, P., Huttenlocher, D.: Pictorial structures for object recognition. International Journal of Computer Vision (IJCV), 55–79 (2005)
14. Singh, V.K., Nevatia, R.: Action recognition in cluttered dynamic scenes using pose-specific part models. In: ICCV (2011)
15. Yang, Y., Ramanan, D.: Articulated pose estimation with flexible mixtures-of-parts. In: CVPR (2011)
16. Sapp, B., Jordan, C., Tasker, B.: Adaptive pose priors for pictorial structures. In: CVPR (2010)
17. Singh, V.K., Nevatia, R., Huang, C.: Efficient Inference with Multiple Heterogeneous Part Detectors for Human Pose Estimation. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part III. LNCS, vol. 6313, pp. 314–327. Springer, Heidelberg (2010)
18. Sapp, B., Toshev, A., Taskar, B.: Cascaded Models for Articulated Pose Estimation. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part II. LNCS, vol. 6312, pp. 406–420. Springer, Heidelberg (2010)
19. Eichner, M., Ferrari, V.: We Are Family: Joint Pose Estimation of Multiple Persons. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part I. LNCS, vol. 6311, pp. 228–242. Springer, Heidelberg (2010)
20. Gu, C., Ren, X.: Discriminative Mixture-of-Templates for Viewpoint Classification. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part V. LNCS, vol. 6315, pp. 408–421. Springer, Heidelberg (2010)
21. Bourdev, L., Malik, J.: Poselets: Body part detectors trained using 3D human pose annotations. In: ICCV (2009)
22. Maire, M., Yu, S.X., Perona, P.: Object detection and segmentation from joint embedding of parts and pixels. In: ICCV (2011)
23. Yang, W., Wang, Y., Mori, G.: Recognizing human actions from still images with latent poses. In: CVPR (2010)
24. Farrell, R., Oza, O., Zhang, N., Morariu, V.I., Darrell, T., Davis, L.S.: Subordinate categorization using volumetric primitives and pose-normalized appearance. In: ICCV (2011)
25. Bourdev, L., Maji, S., Malik, J.: Describing people: Poselet-based approach to attribute classification. In: ICCV (2011)
26. Wang, Y., Tran, D., Liao, Z.: Learning hierarchical poselets for human parsing. In: CVPR (2011)
27. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR (2005)