Detecting People Using Mutually Consistent Poselet Activations *

Lubomir Bourdev[1,2], Subhransu Maji[1], Thomas Brox[1], and Jitendra Malik[1]

[1] University of California at Berkeley
[2] Adobe Systems, Inc., San Jose, CA
{lbourdev,smaji,brox,malik}@eecs.berkeley.edu

Abstract. Bourdev and Malik (ICCV 09) introduced a new notion of parts, poselets, constructed to be tightly clustered both in the configuration space of keypoints, as well as in the appearance space of image patches. In this paper we develop a new algorithm for detecting people using poselets. Unlike that work which used 3D annotations of keypoints, we use only 2D annotations which are much easier for naive human annotators. The main algorithmic contribution is in how we use the pattern of poselet activations. Individual poselet activations are noisy, but considering the spatial context of each can provide vital disambiguating information, just as object detection can be improved by considering the detection scores of nearby objects in the scene. This can be done by training a two-layer feed-forward network with weights set using a max margin technique. The refined poselet activations are then clustered into mutually consistent hypotheses where consistency is based on empirically determined spatial keypoint distributions. Finally, bounding boxes are predicted for each person hypothesis and shape masks are aligned to edges in the image to provide a segmentation. To the best of our knowledge, the resulting system is the current best performer on the task of people detection and segmentation with an average precision of 47.8% and 40.5% respectively on PASCAL VOC 2009.

1 Introduction

Detecting people in images is hard because of the variation in visual appearance caused by changes in clothing, pose, articulation and occlusion. It is widely accepted that a representation based on parts is necessary to tackle the challenge of detecting people in images. But how shall we define parts?

Historically, the most common choice has been to use basic anatomical structures such as torso, left upper arm, left lower arm, and in a probabilistic framework such as pictorial structures [1], these become nodes in a graphical model and the conditional independence assumption inherent in the tree structure make inference tractable. Other approaches that look for good scoring parts in the right spatial relationships may be found in [2–6].

* This work was supported by Adobe Systems, Inc., a Google Fellowship., the German Academic Exchange Service (DAAD), and ONR MURI N00014-06-1-0734.

While these parts are quite natural in constructing kinematic models of a moving person, they are not necessarily the most salient features for visual recognition. A limb, modeled as a pair of parallel line segments, is quite difficult to detect reliably; there are false positives all over an image. In contrast, a visual conjunction such as "half of a frontal face and a left shoulder" may be a perfectly good discriminative visual pattern. This is perhaps the reason why the best performing approaches on people detection tend not to be based on first detecting anatomical parts. Leading this trend was work on pedestrian detection [7, 8] using a multi-scale sliding window paradigm; other examples of such "appearance-based" techniques include [9–11, 4]. Currently the best performing system on the task of people detection is by Felzenszwalb et al. [12] who generalized the approach to allow an intermediate layer of "parts" that can now be shifted with respect to each other, rendering the overall model deformable. The templates for these parts emerge as part of the overall discriminative training. The latest version, dubbed *Latent SVM* by the authors, has an additional mixture model on top permitting a rudimentary treatment of aspect.

Bourdev and Malik [14] introduced a new notion of parts as *poselets*, where the key idea is to define parts that are tightly clustered both in configuration space (as might be parameterized by the locations of various joints), and in appearance space (as might be parameterized by pixel values in an image patch). Finding such parts requires extra annotation, and [14] introduced a new dataset, H3D, consisting of images of people annotated with 3D keypoints making use of Taylor's algorithm [15]. The poselets themselves are created by a search procedure. A patch is randomly chosen in the image of a randomly picked person (the *seed* of the poselet), and other examples are found by searching in images of other people for a patch where the configuration of keypoints is similar to that in the seed (see figures 1, 6, and 7 in [14]). Given a set of examples of a poselet, which are, by construction, tightly clustered in configuration space, HOG features [7] are computed for each of the associated image patches. These are positive examples for training a linear Support Vector Machine. At test time, a multi-scale sliding window paradigm is used to find strong activations of the different poselet filters. These are combined by voting using a Max Margin Hough Transform for the torso/bounding box of a person.

In this paper, we present a better way to define and use poselets. We start with a critique of the approach in [14]:

The use of 3D keypoint annotations: While these carry more information than 2D annotations, they come at a cost in terms of annotation expense. The H3D annotation environment requires some degree of skill, and about 1-2 minutes per image. If we only mark keypoints in 2D, the task becomes much simpler and portable to unskilled labor of the type available on Amazon Mechanical Turk. While individual annotations become less informative, the ability to collect many more for a given amount of time and money is a great advantage. Additionally this makes the poselet idea applicable to other object categories where lifting to 3D using the Taylor algorithm is not even possible.

The use of Hough Transform voting: While such techniques have been used in computer vision from early days, and are natural baselines before trying more complex approaches, they provide less flexibility than one might like. Essentially this is a star model [16], in the graphical model sense, with part positions being referred to a center (a torso in [14]). However there may be no common target that all parts predict reliably. Each poselet makes good predictions only about local structure – a feet poselet does not know if the person is sitting or standing, and a face poselet cannot know if the person is occluded by, say, a table. Instead, we should look at pairwise consistency of poselets. A left shoulder poselet and a frontal face poselet may be uncertain in their prediction of the visible bounds, but they are certain on where the shoulders are, which makes it easier to tell if they refer to the same person.

In the following sections we propose solutions to these limitations which significantly increase performance.

2 Overview of our Approach

The first step is to train poselets using only 2D keypoint annotations. Ignoring 3D information becomes possible by a new distance function for comparing 2D keypoint configurations. This simplification of annotation allowed us to augment the training set of [14] by annotation of the people category of PASCAL VOC 2009 training and validation images. The larger amount of training data leads to better initial poselet detectors. The experiments in this paper are based on training 500 such detectors, and we select from these, in a greedy fashion, the 100 or 200 best performing poselets that maximize coverage of the different examples in the training set.

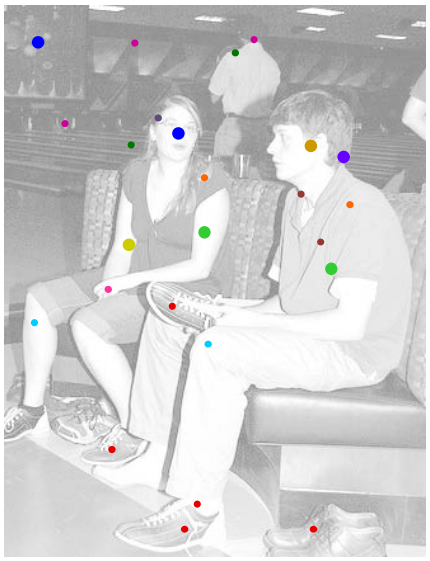
At test time, a multi-scale sliding window paradigm is used to find strong activations of the different poselet filters. In the overview figure for our algorithm, the results of this stage are shown as Fig. 1.1. We need to cluster these activations together if they correspond to the same hypothesized person in the image, predict a score for this person hypothesis, as well as an associated figure/ground segmentation and a bounding box.

The key insight here is that if two poselet activations are consistent, they will make similar predictions of the keypoints of the person, because two consistent true positive activations detect parts of the *same* person.

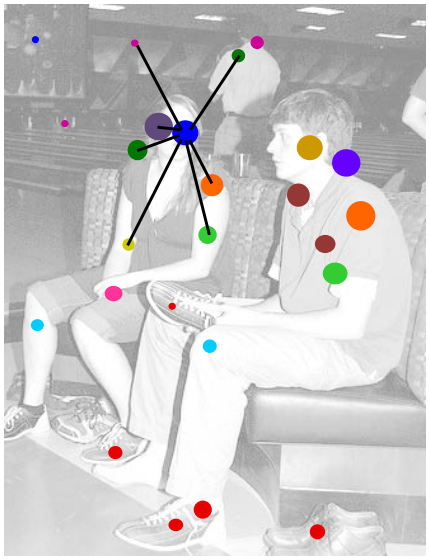
At training time, we can measure the empirical keypoint distributions (Fig. 2) associated with true activations of various poselet types, and at test time, we measure consistency between two poselet activations i and j using the symmetrized KL-divergence of their empirical keypoint distributions $\mathcal{N}_i^k$ and $\mathcal{N}_j^k$:

$$D_{SKL}(\mathcal{N}_i^k, \mathcal{N}_j^k) = D_{KL}(\mathcal{N}_i^k || \mathcal{N}_j^k) + D_{KL}(\mathcal{N}_j^k || \mathcal{N}_i^k) \tag{1}$$

$$d_{i,j} = \frac{1}{K} \sum_k D_{SKL}(\mathcal{N}_i^k, \mathcal{N}_j^k) \tag{2}$$

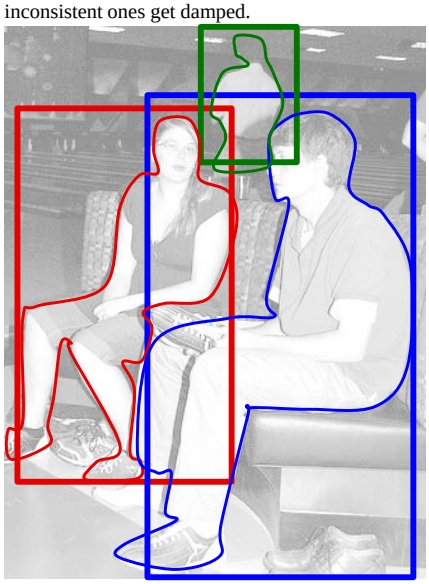


1. q-scores. Different colors illustrate different poselet detectors firing in the image. The blob size illustrates the score of the independent poselet classifier.



2. Q-scores (Section 4). Evidence from consistent poselet activations leads to a reranking based on mutual activation (Q-scores). Weaker activations consistent with others gain importance, whereas inconsistent ones get damped.

3. Clustering (Section 5). Activations are merged in a greedy manner starting with the strongest activation. Merging is based on pairwise consistency.



4. Bounding boxes (Section 6) and segmentations (Section 7). We predict the visible bounds and the contour of the person using the poselets within the cluster.

Fig. 1. Schematic overview with manually marked activations to illustrate the method we propose in this paper.

Fig. 2. Empirical keypoint distribution: locations of the shoulders (left), shoulder and ear (middle), and shoulders and hips (right) over true positive poselet activations.

Since we represent these keypoint distributions as 2D Gaussians, D_{SLK} has a closed-form solution, and the summation is over all the K common keypoints in the two annotations.

The step from Fig. 1.1 to Fig. 1.2 illustrates an additional layer in the detector that uses the context of other poselet activations. This can be regarded as a feed-forward network, where the first layer generates poselet activations whose scores are independent (we call them **q-scores**) and the second layer combines all these to result in context-improved rescoring **Q-scores**. Alternatively, the **q** to **Q** stage can also be regarded as a star model applied to each poselet activation. The number of poselet activations stays the same, but the score of each activation is changed.

The activations are then clustered together to form people detections; cf. Fig. 1.3. We use a saliency based agglomerative clustering with pairwise distances based on consistency of the empirical keypoint distributions predicted by each poselet. Activations that have low score and are not consistent enough to be merged with one of the existing clusters get removed.

Fig. 1.4 illustrates the final step of predicting bounding boxes from the poselets in each cluster. Alternatively, we can predict segmentations from the clustered poselets.

3 Training and Selecting Poselets

We used the H3D training set (750 annotations), the PASCAL VOC 09 training set (2819 annotations for which we added keypoints), and 240 annotations we added manually from Flickr. We doubled this set by mirroring the images horizontally. Our training algorithm consists of the following steps:

1. Collecting patches. We select 500 random windows from the training set (*seed* windows), sampling with uniform distribution over scale and position while keeping a fixed aspect ratio of 1.5. For each seed window we extract patches from other training examples that have similar local keypoint configuration. Following [14], we compute a similarity transform that aligns the keypoints of

each annotated image of a person with the keypoint configuration within the seed window and we discard any annotations whose residual error is too high. In the absence of 3D annotation, we propose the following distance metric:

$$D(P1, P2) = D_{proc}(P1, P2) + \lambda D_{vis}(P1, P2), \tag{3}$$

where D_{proc} is the Procrustes distance between the common keypoints in the seed and destination patch and D_{vis} is a visibility distance, set to the intersection over union of the keypoints present in both patches. D_{vis} has the effect of ensuring that the two configurations have a similar aspect, which is an important cue when 3D information is not available. Note that we allow rotations as part of the similarity transformation during alignment, which helps augment the useful part of the training set for a poselet.

2. Classifier training. We construct HOG features [7] from the collected patches and from random negative example patches and we train linear SVM classifiers. One important difference from [14] is that instead of using all patches as training examples we only use the nearest 250 training examples. Given the size of our training set, this ensures that all the training patches are sufficiently close to the "seed" patch. Otherwise, what may happen is that, e.g., as we collect more examples for a profile face detector, we will eventually start including examples of frontal faces, and they will end up dominating the classifier. Following standard practice, we bootstrap the initially trained SVMs by scanning over images that contain no people, collecting hard false positives and retraining. This process culminates in 500 trained poselet classifiers.

3. Finding true and false positives. We do a scanning pass over our training set and collect the top $2N$ activations of each poselet, where N is the number of annotated people. We assign labels (true positive, false positive, unknown) to each poselet activation. To assign a label we use the bounds of the patches we extracted in step 1. We partition the bounds into two classes: the top-rank patches (the training patches) are treated as ground truth; the lower-rank patches are treated as secondary ground truth. Any activation that has intersection over union overlap of more than 0.35 with a ground truth is assigned a true positive label. If the overlap with a secondary ground truth is less than 0.1 or none, it is assigned a false positive label. All other cases remain unlabeled.

4. Collecting information on each poselet.

(a) We fit a logistic over the positive and negative activations and the associated scores to convert SVM scores into probabilities q_i.
(b) We set a threshold for the SVM score that ensures 90% of the positive and unlabeled examples are above the threshold. This allows each poselet's detection rate to match the frequency of the pattern it has learned to detect.
(c) We fit a model for the keypoint predictions conditioned on each poselet by observing the keypoint distributions of the true positive activations of each poselet type. An example is shown in Fig. 2. We model the distributions using a 2D Gaussian associated with each keypoint.
(d) We fit the prediction of the visible bounds of the human relative to the poselet in a similar way using the true positive activations. We find the mean and variance of x_{min}, y_{min}, x_{max}, y_{max} of the visible bounding box.

5. Poselet selection. The 500 poselet detectors trained in the previous stages are based on randomly selected seed windows, and as a consequence some of these will be redundant and others will correspond to rare patterns. This suggests that we could select a smaller subset that could provide nearly equivalent or even better performance to the whole set (analogous to feature selection for classifiers). We treat this as a "set cover" problem, and solve it using a greedy strategy. For every positive example in the training set, we determine which poselets "cover" it, in the sense that the poselet has an above threshold activation which overlaps it sufficiently (step 3 above). We first pick the poselet that covers the most examples, then incrementally add poselets that cover the most not yet covered examples. Once there is no poselet that can cover any previously uncovered example, we select the poselet that covers the most examples covered by only one previous poselet, etc.

4 Exploiting Context among Poselets

When examining poselet activations, it becomes clear that they are far from perfect. This is but to be expected; the low level signal captured by HOG features is often ambiguous. Sometimes there is just not enough training data, but sometimes there are also "near-metamers"; patterns that can be distinguished by a human observer using additional context, but are almost indistinguishable given the HOG signal inside the image patch. For example, a back-facing head-and-torso pattern is similar in appearance to a front-facing one, and thus a back-facing poselet will often fire on front-facing people as well; see Fig. 3. Another example is a left leg, which in isolation looks very similar to a right leg.

One can resolve these ambiguities by exploiting context – the signal within a patch may be weak, but there is strong signal outside the patch or at a different resolution. We use the pattern of neighboring poselet activations for disambigua-

Fig. 3. The top 10 activations of a poselet trained to find back-facing people. **Top row:** Sorted by q-score. **Bottom row:** Sorted by Q-score. The correct and false activations have a green or red bounding box, respectively. The Q-scores are computed using the context of other activations, e.g. frontal faces, to disambiguate front-facing from back-facing people. Without context we make 6 mistakes (top) whereas using context we make only two mistakes (bottom).

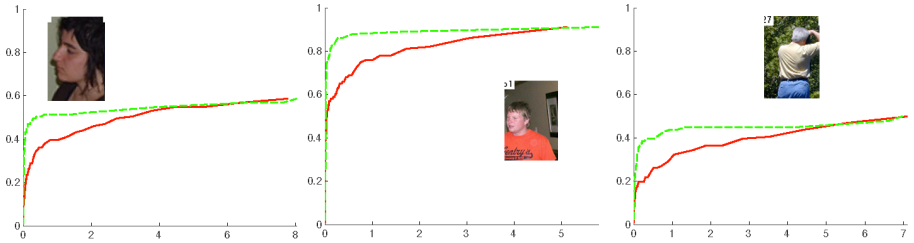


Fig. 4. ROC curves for activations of three poselets computed on our test set. Red continuous lines use q score and green dashed lines use Q score.

tion. For example if a frontal face poselet fires strongly, we can infer that we are more likely to have a front-facing head-and-shoulder pattern, rather than a back-facing one. The oval shape of a wheel sometimes triggers a face detector, but we can suppress the detection if there is no torso underneath.

We refer to the score of a poselet activation based only on its classifier as **q-score** and one that uses other nearby activations as **Q-score**. For each activation i we construct a context feature vector F_i of size the number of poselet types. The pth entry of F_i is the maximum q-score q_j over all activations j of poselet p that are consistent with activation i (or zero if none). We train a linear SVM on the context feature vectors of activations in the training set using their true and false positive labels. We then train a logistic to convert the SVM score into a probability Q_i. The result is what we call Q-score.

We treat two activations i and j as consistent if the symmetrized KL divergence, as defined in (2), $d_{i,j} < \tau$. We set τ as the threshold that best separates distances among consistent activations from distances among inconsistent activations on the training set. For all pairs of labeled activations on the training set we can determine whether they are consistent or not - namely, two activations are consistent if they are both true positives and share the same annotation.

Fig. 3 shows examples of the top activations of our back-facing pedestrian sorted by q-score and below them the corresponding top activations sorted by Q-score. Fig. 4 shows typical ROC cuves with q-scores vs Q-scores for the same poselet. Clearly, the mutual context among activations helps to obtain a better ranking. It is worth noting that Q-scores are assigned to the same activations as the q-scores. While the ranking is changed, the localization of the activation stays the same.

5 Clustering Poselet Activations

Our earlier approach in [14] is build upon the Max Margin Hough Transform from [17] in order to group poselet activations to consistent people detections. This comes with the assumption that the object has a stable central part and the relative position of all other parts has very small variance – an assumption that is not satisfied for articulated objects, such as people. We propose an alternative clustering algorithm:

Fig. 5. Examples of poselet activations during clustering. The activation bounding boxes and the predictions of the hips and shoulders are shown. **Left:** We start with the highest probability activation, which for this image is a left shoulder. **Center:** Example of two compatible activations which will be placed in the same cluster. **Right:** Example of incompatible activations which will end up in separate clusters.

1. Initialize the set of clusters that correspond to person detection hypotheses $M = \{\emptyset\}$.
2. Successively take the poselet activation a_i with the highest score Q_i:
   (a) Find the closest cluster $m_j = \text{argmin}_{m_j \in M}\, d(a_i, m_j)$, where the distance d from a_i to cluster m_j is estimated using average linkage.
   (b) If $d(a_i, m_j) < \tau$ then $m_j \leftarrow merge(m_j, a_i)$, i.e. we merge i into an existing cluster. Otherwise, if $|M| < t$ then $M \leftarrow \{M \cup a_i\}$, i.e. we form a new cluster.

In the end the poselet activations are grouped into clusters each corresponding to a person detection hypothesis. In addition some poselets with low scores that are inconsistent with any clusters are marked as false positives and are discarded. The parameter t is a tradeoff between speed and false positive rate. We set $t = 100$, i.e. we collect at most 100 person hypotheses from each image.

This algorithm is a form of greedy clustering starting with the highest- probability poselet activations. Compared to other schemes such as spectral clustering or agglomerative clustering, the proposed algorithm has computational advantages because it processes the most salient information first. The algorithm runs in linear time. We do not spend compute cycles measuring distances between low scoring detections, and the algorithm can be terminated at any time with a good list of the most-salient-so-far hypothesis M. Furthermore, by starting with the highest probability detections we are less likely to be mislead by false positives. Fig. 5 shows examples of merging compatible activations (center) and forming a new cluster (right).

6 Locating and Scoring People Hypotheses

Given a cluster of poselet activations, we can predict the location of the torso, as well as a visible bounding box. We can also compute a score S, which is a measure of how likely the cluster corresponds to a person as opposed to being a false positive.

1. Torso prediction. The human torso is a more stable region to predict than the visible bounding box. Thus, before applying non-maximum suppression, we first predict torsos and derive visible bounds from these predictions. The torso can be predicted from the poselet activations within a cluster. If we use context, we also include all compatible activations that might not be in the cluster. We predict the locations of the hips and shoulders as the average prediction of each poselet activation, weighted by the score of the activation. These four keypoints define the torso of the person, which we parameterize using (x,y) location, length and angle. We use a fixed aspect ratio of 1.5.
2. Non-maximum suppression. We use agglomerative clustering to merge clusters whose intersection-over-union of torso bounds is greater than 0.6.
3. Visible bounds prediction. For each activation in the merged clusters we compute its prediction for the expected visible bounds x_{min}, y_{min}, x_{max} and y_{max} and the associated variances. We then perform mean shift for each of the four estimates independently and pick the dominant mode. Mean shift allows us to take into account the variance of the prediction, which is important. A frontal face poselet, for example, has a very reliable prediction for y_{min}, but is very unreliable for y_{max} since sometimes the legs of the person may be occluded.
4. Improving the predicted bounds. The above generative bounding box prediction is not very accurate and we enhance it using a linear regression similar to [12]. Specifically we transform $[x_{min}y_{min}x_{max}y_{max}]$ with a 4x4 regression matrix T. To train T, we perform steps 1, 2, and 3 on the training set, we match the bounds predictions to the ground truths using intersection over union overlap of 0.45 and collect the true positives. We then fit T using the predicted bounds and the associated ground truth bounds.
5. Computing the score of a poselet cluster. We follow [14] to predict the score S of the poselet cluster, i.e., we train a linear discriminative classifier with positivity constraints on its weights to predict the scores based on the activations within the cluster. We can use q-scores or Q-scores here, and we will show a comparison in Section 8. For our positive examples we use detections on the training set whose bounds intersection over union overlap is over 0.5. For negative examples we use detections that do not intersect the truth or whose overlap is less than 0.1. Our feature vector has the dimensionality of the number of poselet types. The feature value for each poselet type is the maximum of all activations of that poselet type within the cluster.

7 Object Segmentation by Contour Alignment

While prediction of bounding boxes is a reasonable proxy for the object detection problem, the final localization task is actually the segmentation of the detected object. From training examples with segmentation masks available we can derive a figure/ground predictor for each poselet. We use a simple shape model for each poselet by just averaging the masks of all examples in the training images after keypoint alignment.

At test time, we can derive a shape prior for the object by integrating the mask predictions $\phi_i : \mathbb{R}^2 \to [0,1]$ of all poselet activations $i = 1, ..., n$ assigned to one cluster. Rather than just averaging the masks, we weight them by the activation scores Q_i

$$p_{\text{in}}(x,y) = \frac{\sum_{i=1}^{n} Q_i \chi_i(x,y) \phi_i(x,y)}{\sum_{i=1}^{n} \chi_i(x,y)}, \tag{4}$$

where $\chi_i : \mathbb{R}^2 \to \{0,1\}$ denotes the indicator function for the poselet's support in the image. As we are interested in a binary segmentation, the soft masks are thresholded at $\theta_m = 0.07$. This value has been optimized for the PASCAL VOC 09 validation set.

The above procedure yields an a priori decision on which pixels belong to an object given the detection of certain poselets. It so far ignores further indication from the image. In order to get a more precise localization of object boundaries we align them to contours in the image. We use the state-of-the-art boundary predictor from [18] to obtain an edge map $f : \mathbb{R}^2 \to [0,1]$ of the image. Moreover, we extract the silhouette $g : \mathbb{R}^2 \to \{0,1\}$ of the predicted binary mask. We then estimate the deformation field $(u,v) : \mathbb{R}^2 \to \mathbb{R}$ that minimizes

$$E(u,v) = \int_{\mathbb{R}^2} |f(x,y) - g(x+u, y+v)| + \alpha \left(|\nabla u|^2 + |\nabla v|^2 \right) dxdy. \tag{5}$$

The parameter $\alpha = 50$ determines the amount of flexibility granted to the deformation field. We use a coarse-to-fine numerical scheme known from optical flow estimation to compute the minimizer of (5) [19]. Warping the initial binary mask with the optimum deformation field (u,v) yields a mask that is aligned with boundaries in the image.

For segmenting the whole image, we paste the aligned binary masks from all clusters into the image domain, ignoring clusters with an overall score $S \leq 12$. Since we run the segmentation for only one category, the ordering of the single detections has no effect.

8 Experiments

Table 1 investigates the effect of the amount of poselets showing results using 10, 40, 100, and 200 selected poselets. Clearly, more poselets first help improving the detection performance, but the improvement saturates between 100 and 200 poselets. Table 1 also shows the positive effect of exploiting mutual context between poselet activations. The AP with Q-scores is consistently larger.

As an important baseline comparison to our previous detection in [14], we evaluated our new detector, using 200 poselets, on the task of detecting human torsos on the H3D test set. The ROC curves are shown in Fig. 6. The new ROC curve outperforms the one from [14] over the entire range. In [14] we required 256 poselets and we also scanned the horizontally flipped version of the image, which has the effect of doubling the poselets to 512.

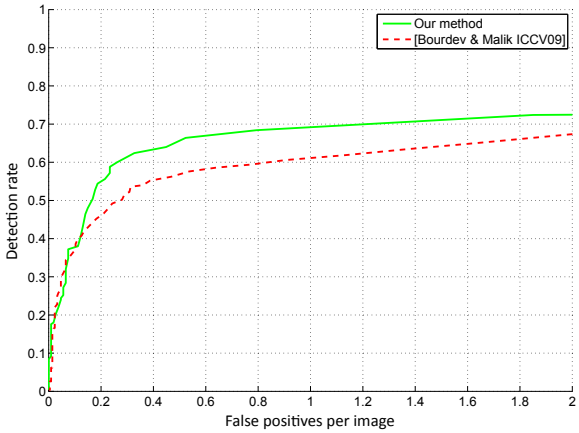


Fig. 6. Performance on the human torso detection task on the H3D test set.

Num. poselets	q-scores	Q-scores
10	36.9%	37.8%
40	43.7%	44.3%
100	45.3%	45.6%
200	45.7%	46.9%

Table 1. AP on PASCAL VOC 2007 test set for various numbers of poselets using q-scores or Q-scores as described in Section 4.

	Detection				Segmentation		
	100 poselets	200 poselets	[12]	[13]	masks only	alignment	[20]
VOC 2007	45.6%	46.9%	36.8%	43.2%			
VOC 2008	54.1%	52.6%	43.1%		41.9%	43.1%	41.3%
VOC 2009	47.8%	46.9%		43.8%	39.4%	40.5%	38.9%

Table 2. Our performance on PASCAL VOC compared to the currently best results reported for the detection and segmentation tasks on the person category. The segmentation results were produced with 200 poselets.

Finally we provide results on the person category of the recent PASCAL VOC challenges. As reported in Table 2, we have the best results reported to date, both in the detection and the segmentation challenge. Our results are reported for the competitions 4 and 6 because our method requires 2D keypoint annotations.

Table 2 also shows the impact of aligning the mask predictions to boundaries in the image. It is relatively small in quantity, as the performance is mainly due to the detector. The visual effect is much larger, as segmentations align well with true object boundaries. Some example detections and segmentations are shown in Fig. 7 and Fig. 8.

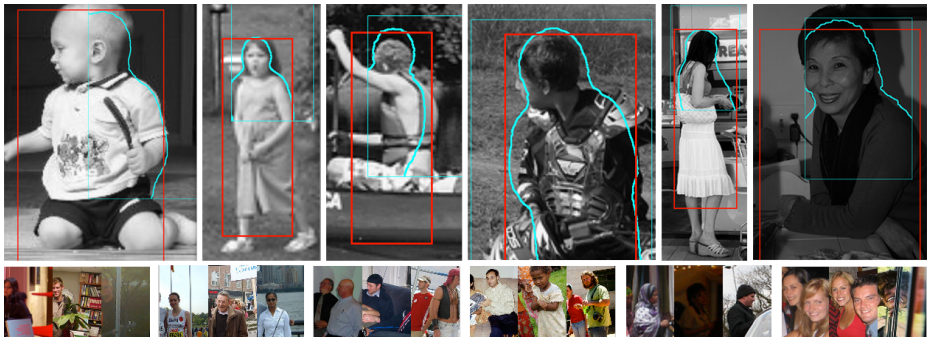


Fig. 7. Detection examples. The person's bounding box is shown in red. The highest probability poselet activation is shown in a cyan bounding box and a figure/ground outline. Below each image we show three training examples from the activated poselet.

Fig. 8. Segmentation examples. The top middle example shows a typical limitation in case of occlusion by other objects.

9 Conclusion

It is possible to view the poselets approach in a natural sequence of increasing complexity from (1) Dalal and Triggs' [7] single holistic model to (2) Felzenszwalb et al.'s [12] parametric part model on to (3) poselets. In [12], certain fixed choices are made: one root filter, six part filters, two components. The poselet framework can be thought of as being in the spirit of nonparametric statistics – models with greater flexibility which, as more training data becomes available, are expected to have superior performance. These performance improvements do not come at an inordinate expense in terms of running time. On a 3GHz Macbook Pro our Matlab implementation with 40 poselets runs in about 27 seconds per image,

where a large part of the time is spent for HOG computation. We conclude by noting that the approach described in this paper for detecting people is equally applicable to other object categories. This is the subject of ongoing research.

References

1. Felzenszwalb, P.F., Huttenlocher, D.P.: Pictorial structures for object recognition. IJCV **61** (2005) 55–79
2. Ren, X., Berg, A.C., Malik, J.: Recovering human body configurations using pairwise constraints between parts. ICCV (2005) 824–831
3. Ramanan, D.: Learning to parse images of articulated bodies. NIPS (2006)
4. Ferrari, V., Marin-Jimenez, M., Zisserman, A.: Progressive search space reduction for human pose estimation. CVPR (2008)
5. Andriluka, M., Roth, S., Schiele, B.: Pictorial structures revisited: People detection and articulated pose estimation. CVPR (2009)
6. Bergtholdt, M., Kappes, J., Schmidt, S., Schnörr, C.: A study of parts-based object class detection using complete graphs. IJCV **87** (2010) 93–117
7. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. CVPR (2005) 886–893
8. Oren, M., Papageorgiou, C., Sinha, P., Osuna, E., Poggio, T.: Pedestrian detection using wavelet templates. CVPR (1997) 193–199
9. Mori, G., Malik, J.: Estimating human body configurations using shape context matching. ECCV (2002) 666–680
10. Leibe, B., Leonardis, A., Schiele, B.: Combined object categorization and segmentation with an implicit shape model. ECCV workshop on statistical learning in computer vision. (2004) 17–32
11. Gavrila, D.M.: A Bayesian, exemplar-based approach to hierarchical shape matching. PAMI **29** (2007) 1408–1421
12. Felzenszwalb, P., Girshick, R., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. PAMI, published online (2009)
13. Felzenszwalb, P., Girshick, R., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. Project website: people.cs.uchicago.edu/∼pff/latent (2010)
14. Bourdev, L., Malik, J.: Poselets: Body part detectors trained using 3d human pose annotations. ICCV (2009)
15. Taylor, C.: Reconstruction of articulated objects from point correspondences in a single uncalibrated image. CVIU **80** (2000) 349–363
16. Crandall, D., Felzenszwalb, P., Huttenlocher, D.: Spatial priors for part-based recognition using statistical models. CVPR (2005) 10–17
17. Maji, S., Malik, J.: Object detection using a max-margin hough tranform. CVPR (2009)
18. Arbelaez, P., Maire, M., Fowlkes, C., Malik, J.: From contours to regions: an empirical evaluation. ICCV (2009)
19. Brox, T., Bruhn, A., Papenberg, N., Weickert, J.: High accuracy optical flow estimation based on a theory for warping. ECCV (2004) 25–36
20. Yang, Y., Hallman, S., Ramanan, D., Fowlkes, C.: Layered object detection for multi-class segmentation. CVPR (2010)