Poselets: Body Part Detectors Trained Using 3D Human Pose Annotations *

Lubomir Bourdev[1,2] and Jitendra Malik[1]
[1]EECS, U.C. Berkeley, Berkeley, CA 94720
[2]Adobe Systems, Inc., 345 Park Ave, San Jose, CA 95110
{lbourdev,malik}@eecs.berkeley.edu

Abstract

We address the classic problems of detection, segmentation and pose estimation of people in images with a novel definition of a part, a poselet. *We postulate two criteria (1) It should be easy to find a poselet given an input image (2) it should be easy to localize the 3D configuration of the person conditioned on the detection of a poselet. To permit this we have built a new dataset, H3D, of annotations of humans in 2D photographs with 3D joint information, inferred using anthropometric constraints. This enables us to implement a data-driven search procedure for finding poselets that are tightly clustered in both 3D joint configuration space as well as 2D image appearance. The algorithm discovers poselets that correspond to frontal and profile faces, pedestrians, head and shoulder views, among others.*

Each poselet provides examples for training a linear SVM classifier which can then be run over the image in a multiscale scanning mode. The outputs of these poselet detectors can be thought of as an intermediate layer of nodes, on top of which one can run a second layer of classification or regression. We show how this permits detection and localization of torsos or keypoints such as left shoulder, nose, etc. Experimental results show that we obtain state of the art performance on people detection in the PASCAL VOC 2007 challenge, among other datasets. We are making publicly available both the H3D dataset as well as the poselet parameters for use by other researchers.

1. Introduction

The Oxford English Dictionary defines *pose* as:

> An attitude or posture of the body, or of a part of the body, esp. one deliberately assumed, or in which a figure is placed for effect, or for artistic purposes.

This definition captures the two aspects of a pose:

Figure 1. Poselets are parts that are tightly clustered in both appearance and configuration space. The figure shows positive examples for some of our poselets: #1 (frontal face), #114 (right arm crossing torso), #20 (pedestrian), #79 (right profile and shoulder) and #138 (legs frontal view).

1. A configuration of body parts such as head, torso, arms and legs arranged in 3D space.
2. The resulting appearance, a 2D image created for a viewer, or a camera.

The *configuration space* of an articulated body can be parameterized by the 3D coordinates of the joints, and the *appearance space* by the pixel values. The configuration space of joints has many degrees of freedom, and the appearance is additionally a function of clothing, illumination and occlusion. These phenomena collectively make the detection, joint localization and segmentation of people in images some of the most challenging problems in computer

*This work was supported by Adobe Systems, Inc., a grant from Hewlett Packard and the MICRO program, as well as ONR MURI N00014-06-1-0734

2009 IEEE 12th International Conference on Computer Vision (ICCV)
978-1-4244-4419-9/09/$25.00 ©2009 IEEE

vision. In keeping with the "divide and conquer" principle, it is natural to think that "parts" could help by factoring the complexity. But how do we decide what good parts are?

The principal contribution of this paper is a new notion of part, a "poselet", and an algorithm for selecting good poselets (Figure 1). We use the term *poselet* to suggest that it describes a part of one's pose. Since we need parts to provide a bridge from appearance space to configuration space, we argue that a "good" poselet must satisfy the following two criteria:

1. It should be easy to find the poselet given the input image. This suggests that the poselet must be tightly clustered in appearance space, because low in-class variability leads to better detection performance for a classifier.
2. It should be easy to localize the 3D configuration of the person conditioned on the detection of a poselet. If a poselet corresponds to a tight cluster in configuration space, then this will be the case, at least over the support of the poselet.

In order to operationalize this intuition that a poselet should be tightly clustered in both appearance and configuration space, we will need training data that is annotated with 3D configuration information, not just a bounding box or the pixel support map for a person. We have developed a novel dataset for this purpose, which we call *H3D*, or *Humans in 3D*. This currently consists of 2000 annotations of humans, including the 3D locations of 19 keypoints (joints, eyes, ears and nose) and 15 kinds of pixel-level labels of image patches, such as "face","hair","upper clothes","left arm", etc. H3D allows us to use the 3D configuration proximity (criterion 2) as a starting point of poselet selection. As we show on Figure 7, that allows us to generate training examples that may vary in appearance but have similar semantics - human heads from behind, or people with crossed hands, or the legs of a pedestrian making a step, etc. As a result, our poselet classifiers are directly trained to handle the visual variation associated with a common *underlying semantics*.

Given a set of poselets (256 linear support vector machines in the current implementation), we scan the input image at multiple scales and use the outputs of these to, in turn, vote for location of the torso bounds or body keypoints (Section 5). This is essentially a Hough transform step in which we weigh each vote using weights learned in a max-margin framework [10].

We can contextualize our research by noting that past work on analyzing images of people has defined parts by considering just one of the two criteria we outlined above:

1. Work in the pictorial structure tradition, from Felzenszwalb and Huttenlocher [6] and others [15, 14, 7, 1], picks a natural definition of part in the 3D configuration space of the body, guided by human anatomy. Even earlier work with "stick figure" representations using generalized cylinders to model various body parts made essentially the same choice [12, 17]. While these parts are the most natural if we want to construct kinematic simulations of a moving person, they may not correspond to the most salient features for visual recognition. It may be that "half of a frontal face and a left shoulder" or "the legs of a person making a step in a profile view" are particularly discriminative visual patterns for detecting a human–does it matter that these are not "parts" in an anatomical sense, or that English doesn't have single words for them? In our context, this line of research satisfies criterion 2, but not criterion 1.

2. Work in the appearance based window classification tradition directly tries to find weights on various features for best classification performance. For example, Oren *et al*. [13] and Dalal and Triggs [3] train holistic classifiers for pedestrian detection. These approaches degrade in the presence of articulations, and Felzenszwalb, McAllester and Ramanan [5] have 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. Such approaches, however, are not suitable for pose extraction or localization of the anatomical body parts or joints. An alternative way to provide flexibility is by the use of point descriptors as in the work of Mori and Malik [11], or Leibe *et al*. [8]. What is common to all these approaches is the parts or point descriptors are chosen based purely on appearance (criterion 1) but not configuration (criterion 2).

Finally, there are now some hybrid approaches which have stages of one type followed by a stage of another type. Ferrari *et al*. [7] start with holistic upper-body detection based purely on appearance, followed by the application of a pictorial structure model in regions of interest. Andriluka *et al*. [1] train part detectors for anatomically defined body parts which then are combined using pictorial structures. Unlike what we propose with poselets, the parts themselves are not jointly optimized with respect to combined appearance and configuration space criteria.

Our method combines the benefits of both of these prior directions of research. We show state-of the-art performance (Section 6), including on the PASCAL VOC 2007 challenge, where our AP for the person category is 0.365. As far as we know the current best result is [5] with AP of 0.368.

2. H3D, a Dataset of Humans in 3D

Human detection and recognition has been a much studied subject and there are many datasets with annotations varying from bounding boxes[1] [4], to region annotations [16, 21] and 2D joint locations [7, 21, 19]. Various data collection schemes have been explored, including internet-scale collaboration [16] and using Mechanical Turk [19] and multiple strategies have been employed for increasing the quality of the data. However, we believe our dataset is unique in cross-referencing the full 3D pose, keypoint visibility and region annotations.[2] We believe 3D information and visibility are very important for generating better training data, building accurate statistics of 3D poses, decomposing camera parameters and many other tasks. Our part selection method will not work well using 2D annotations.

H3D currently consists of 2000 annotations[3] which we have split into 1500 training 500 test human annotations. We have chosen the images from Flickr with Creative Commons Attributions License[4] which allows free redistribution and derivative work. H3D provides annotation of 15 types of regions of a person (such as "face", "upper clothes", "hair", "hat", "left leg", "background") and 19 types of keypoint annotations, which include joints, eyes, nose, etc. Cross-referencing appearance and 3D structure allows us to do new and powerful types of queries for pose statistics and appearance, described below. The dataset is available on our web site.

Our annotation environment is shown on Figure 2. The time to create an annotation varies on the difficulty of the annotation and the expertise of the annotator, but on average it takes about 5 minutes to specify the keypoints, set the 3D pose and label the regions. In this section we give examples of some of the types of queries supported by H3D.

- **Keypoint Distributions:** Using the annotated keypoint locations, we can determine the expected image locations of a set of keypoints conditioned on the locations of other keypoints. Such distributions would be valuable for pose extraction algorithms. Figure 5 (right) shows our prediction for the locations of the left ankle and right elbow conditioned on the shoulder locations for frontal views. We generated these by transforming each 3D pose to match the shoulder locations and then plotting the projections of the left ankle and right elbow. Note that H3D leverages the data much more than traditional labelled 2D image datasets. The same data could also be generated using a traditional dataset of images with 2D annotated joints, but only images of frontal view annotations will contribute to the statistics. In contrast, H3D projects every 3D pose to the desired view and thus every annotation will contribute to the statistics.

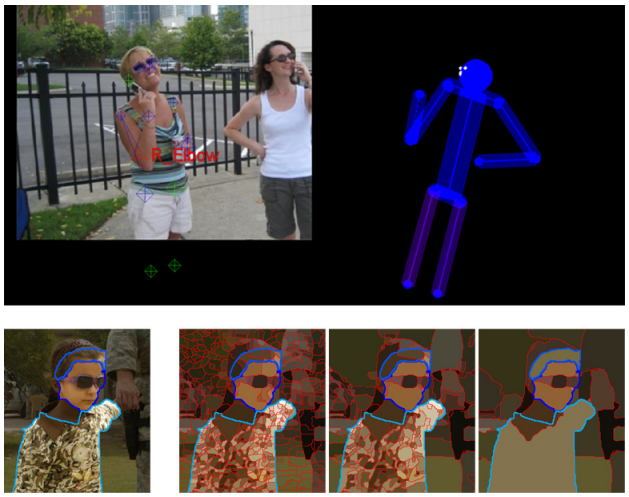

Figure 2. **Top:** Our Java3D annotation tool allows the user to mark keypoints and displays the 3D pose in real time. Users can mark the image while looking at the 3D pose from another view. Our 3D extraction is based on the Taylor method [20] which we have extended to further ease accurate reconstruction. We have introduced extra constraints and global controls that use gradient descent in the overconstrained space to help the user adjust the pose, straighten the torso, etc. **Bottom:** Our region labelling tool performs hierarchical oversegmentation of the image using [9] to allow the user to efficiently and accurately assign region labels. Users start labelling a rough version and refine the labels.

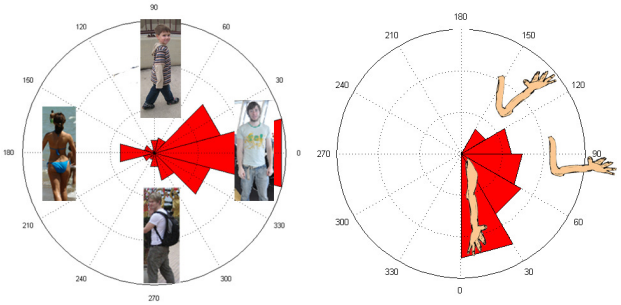

Figure 3. H3D can decompose camera view point, pose and appearance and allow us to model them separately. **Left:** Camera azimuth angle relative to frontal pose. In our dataset 39.6% of the time the human pose is frontal (between -15° and 15°) **Right:** Expected arm bending angle. 33% of the time the arm is almost straight, bent less than 30% degrees. The physical constraints of the human body are implicitly captured in the H3D statistics: no arms bend backwards.

[1] http://pascal.inrialpes.fr/data/human

[2] Motion capture datasets like HumanEva [18] do provide 3D information, but they do not capture the statistics of "people in the wild", neither in terms of pose nor appearance. We are interested in the distribution of poses and appearances of people in typical consumer photo albums.

[3] There are 1000 real annotations, which we have doubled by mirroring them along a vertical axis. Most H3D tasks have no axial symmetry, so the mirrored versions are just as important. We also ensure that there are no images in the training set whose mirrored version is in the test set

[4] http://creativecommons.org/licenses/by/3.0

- **3D Pose Statistics:** The statistics in 2D are less smooth than in 3D because of foreshortening. For example, the left ankle can approach the shoulder in Figure 5 (right) if the person is lying down with legs towards the camera. On the other hand, in 3D the statistics are smoother due to physical length constraints. Since we have the relative 3D coordinates of all joints, we can compute expected 3D joint locations. In Figure 3 (right) we have explored the distribution of angles between the upper and lower arm segments.
- **Camera View Statistics:** Figure 3 (left) also shows that H3D can decompose camera view point from pose and produce separate statistics for each.
- **Appearance Queries:** Registering 3D views with 2D images is powerful, as it allows us to query for the appearance of poselets. Given the normalized locations of two keypoints (which define a similarity transform), a target aspect ratio and resolution, H3D can extract patches from the annotated images. H3D can also leverage our region annotations to include or exclude specific regions. Figure 4 shows the result of displaying people whose hip-to-torso angle is less than 130 degrees (i.e. sitting people). We show them with the background and any occluders masked out.
- **Pixel Label Probability Statistics:** Figure 5 (left) shows how H3D can generate region probability masks conditioned on joint locations and pose. Traditional 2D datasets could also be used to generate such masks, but they will be noisy due to foreshortening. For example, a closeup of a person turned at 45 degrees will have the same keypoint projection of the eyes/hips than a far away person facing the camera, but the spatial probability of their upper clothes regions will be very different due to scale, and without 3D information we cannot distinguish between the two. We have used H3D's ability to generate soft region labels as shown on Figure 12.

Figure 4. Sitting people with the background masked out. To select sitting people, we asked H3D for people whose torso-to-hip angle is below 130°. They are shown sorted by torso-to-hip angle.

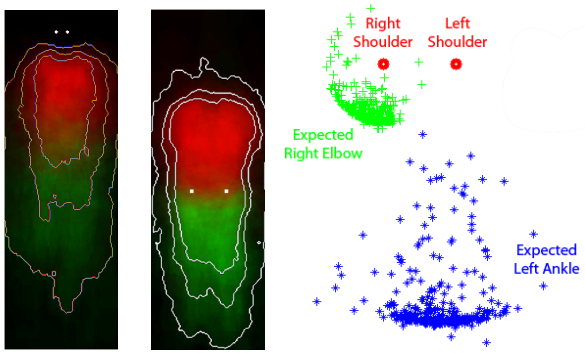

Figure 5. **Left:**H3D can generate conditional region probability masks. Here we show the probability of upper clothes (red) and lower clothes (green) given the location of the eyes (left picture) or hips (right picture) of frontal-facing people. Contours at 0.1, 0.3 and 0.5 are shown. We could compute conditional probabilities on more variables if the size of our dataset allows for meaningful predictions. **Right:**H3D can generate scatter plots of the 2D screen locations of the right elbow and left ankle given the locations of both shoulders. Because the data is in 3D, it leverages information even from profile-facing poses by turning them around and projecting their keypoints.

3. Finding Poselet Candidates

Figure 6 illustrates what we mean by a poselet. A *poselet* describes a particular part of the human pose under a given viewpoint. It is defined with a set of *examples* that are close in 3D configuration space. It is useful to think of poselets as clusters in configuration space. Each example of a poselet has a corresponding rectangular patch at a given position, orientation and scale from the annotation of a given person (bottom row of Figure 6). Each example also corresponds to a point in the configuration space of 3D poses. It is defined as the coordinates of keypoints of the human pose transformed into the example's coordinate space (top row of Figure 6). We define the (asymmetric) distance in configuration space from example s to example r as:

$$d_s(r) = \sum_i w_s(i)||\mathbf{x_s}(i) - \mathbf{x_r}(i)||_2^2(1 + h_{s,r}(i)) \quad (1)$$

where $\mathbf{x_s}(i) = [x, y, z]$ are the normalized 3D coordinates of the i-th keypoint of the example s. The weight term $w_s(i) \propto exp(-\mathbf{x_s}(i)^2/(2\sigma^2))$ is a Gaussian with mean at the center of the patch. The Gaussian is designed to give high weights to terms near the center and low weights to terms far outside the patch. (While the top row of Figure 6 only shows keypoints inside the patch, we also consider nearby keypoints outside the patch). The term $h_{s,r}(i)$ is a penalty based on the visibility mismatch of keypoint i in the two examples. If keypoint i is visible or invisible in both examples, then $h_{s,r}(i) = 0$. Otherwise $h_{s,r}(i) = a, a > 0$.

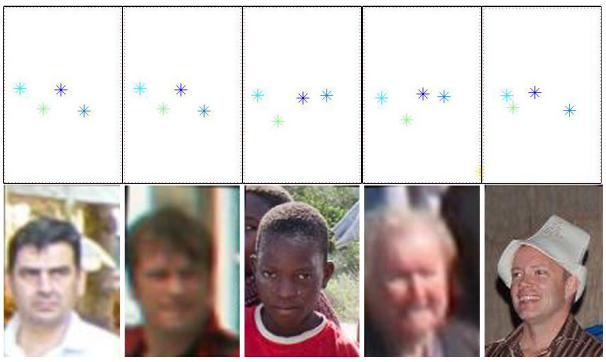

Figure 6. A poselet describing a frontal face and five of its examples. **Top row:** The configuration spaces showing the eyes, nose and left ear keypoints. **Bottom row:** the corresponding image patches. By construction all examples of a poselet have similar configurations and are therefore semantically similar.

Figure 7. Example query regions (left column) and the corresponding closest matches in configuration space generated by H3D. Configuration space proximity tends to produce *semantically* similar examples, although they may be visually very different. The first row, for example, tends to generate frontal-facing people whose left hand is raised near their head. The second row shows examples whose right foot is closer to the camera than their left foot; i.e. matching is done in 3D space.

In addition, it is possible that the i-th keypoint be present in one example but missing from the other. In this case the respective term is $w_s(i)b$ where (σ, a, b, h) are fixed parameters of the model.

Given an example s from human annotation a_s, H3D can efficiently find the corresponding closest example r from annotation a_r. In particular, H3D uses a weighted least squares fit based on $d_s(r)$ to derive the similarity transform (t_x, t_y, α, s) that brings the keypoints of annotation a_r as close as possible to the normalized coordinates of s. We can then measure the quality of the match based on the residual distance $d_s(r)$. Note that the distance is measured in 3D space which allows us to distinguish between keypoints near each other and ones with large foreshortening and learn them in different poselets. Figure 7 shows some query examples on the left and their corresponding closest matches on the right. Notice how our pose space proximity results in examples that, while visually different, are semantically quite similar. This is a very important advantage of our method: our poselet classifiers are going to learn the kind of visual dissimilarity that corresponds to instances of the same semantic class, and thus learn to recognize the semantic class.

We have a simple and efficient procedure to generate a poselet candidate from our training data: Given a rectangular window from one human annotation, we use the above described least-squares method to find the closest corresponding window from every other human annotation in our training set and we keep the examples whose residual distance is less than λ. The parameter λ controls the tradeoff between quantity and quality of the examples. For instance, for a very aggressive setting our frontal face poselet will start to include some profile faces as well. We set λ empirically to a value of 0.1 which results in lots of examples without affecting too much the quality.

Using the above procedure we could generate hundreds of thousands of poselet candidates, for example by starting from random windows. We chose instead to run a scanning window over all positions and scales of all annotations in our training set. We don't need to search over orientation as our least-squares fit will discover rotated examples of the same poselet. This procedure results in about 120K poselets, which, by construction, are semantically tight. We then prune them by removing poselets with very few examples (which correspond to rare configurations) and poselets that are too close to each other in configuration space (which could happen as a result of double-counting during scanning)[5]. This left us with about 2000 poselet candidates.

4. Selecting and Training Poselets

We train classifiers to detect the presence of each poselet by using the examples of the poselet as positive examples, and random image patches from images not containing people as the negative examples. We use a linear SVM and our features are Histograms of Oriented Gradients as proposed by Dalal and Triggs [3]. We use their recommended settings for all parameters, except our scan window has dimensions of 96x64. We train using bootstrapping: we train an initial classifier using the positive and a random set of negative examples, then we use it to scan over images not containing people and collect false positives, and then we do a second round of training by including these hard false positives into the negative training set.

Not all 2000 poselet candidates are suitable for training – some may not train well and others may be redundant. To reduce the computational complexity, we first prune the set of poselet candidates by an order of magnitude: Using an estimate of their cross-validation score and their pairwise

[5] In our current implementation we do not scan over rotations; thus we also remove poselets that have wide orientation variance

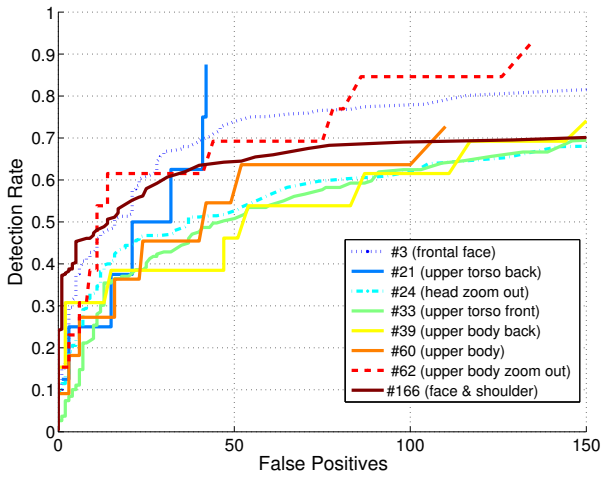

Figure 8. The comparative performance of our top poselets in isolation evaluated on the H3D test set. Poselets focused on the upper body tend to perform best.

distances in configuration space, we use a greedy search to choose a subset of 300 poselets that have high cross validation score and are not too close in configuration space to other chosen poselets. We also filter their examples by pruning examples that the trained SVM scores lower than a given threshold η. Such examples typically include severe occlusions and removing them improves performance.[6].

Examples of some of our poselets are shown on Figure 1. Our algorithm selects frontal faces and profile faces at different zooms, people facing backwards, pedestrians, upper bodies, frontal torso detectors, legs, etc. A lot of the poselets cover the face and shoulder areas, which suggests that those regions are important for detecting people. The head and a shoulder were visible in almost all of the annotated people. Figure 8 shows the performance of our top poselets in isolation. The ROC curve is obtained by running the poselet detectors over the test set and using the H3D distance (Formula 1) to distinguish true from false positives. The top performers are poselets focused on the face at various zoom levels. We believe this is because the frontal face is both easy to detect and occurring in many images of people. Legs, on the other hand, are often occluded whereas hands have large rotation variability and are thus harder to detect.

5. Combining Poselets for Detection and Localization

By construction a poselet is tightly clustered in configuration space, which makes it effective for estimating the local configuration. We explore this property of poselets for detecting torsos and for localizing keypoints of the human body. In this section we refer to torso bounds and keypoints as simply "objects". We use the Generalized Hough Transform framework. Using the H3D training set we fit the transformation from the poselet location to the object. We run each poselet detector at every position and scale of the input image, collect all hits and use mean shift to cluster nearby hits. Each cluster casts a vote for the object location. The probability of detecting the object O at position x is:

$$P(O|x) \propto \sum_i w_i a_i(x) \tag{2}$$

where $a_i(x)$ is the score that a poselet classifier assigns to location x and w_i is the weight of the poselet. To find the peaks in Hough space we cluster the cast votes using agglomerative clustering and we compute the sum in Formula 2 over the poselets within each cluster.

The weights w are used to account for the fact that (1) some poselets are more discriminative than others and (2) the responses of some subsets of poselets are redundant. Note also that we train separate weights for each task as weights are task-dependent. For example, a frontal face poselet is more discriminating for detecting an eye than for detecting an elbow.

We use the Max Margin Hough Transform (M^2HT) proposed by Maji and Malik [10] to learn the weights. Intuitively, some peaks in Hough space correspond to true detections and others are false positives. M^2HT is a discriminative technique that finds the set of weights that maximize the true positive peaks and minimize the false positives. To train for the weights, for each detection task we compute the Hough peaks in the training set (using Formula 2 with $w_i = 1$) and then we find w using the following optimization problem:

$$\min_{w,b,\xi} \frac{1}{2} w^T w + C \sum_{i=1}^{T} \xi_i \tag{3}$$

$$\text{s.t. } y_i(w^T A_i + b) \geq 1 - \xi_i \tag{4}$$

$$w \geq 0, \xi_i \geq 0, \forall i = 1, 2, ..., N \tag{5}$$

where A_i^j is the score of poselet j in Hough peak i (or 0 if the poselet did not vote for the peak) and $y_i = 1$ if the peak is true positive and -1 if false positive. The formulation is similar to an SVM with the additional positivity constraint on the weights.

6. Experimental Results

Detecting Human Torsos. Figure 9 shows the performance of our torso detector together with other published detectors on the H3D test set. We used the PASCAL VOC criterion [4] for overlap to determine true from false positives

[6]Since we cannot afford to do elaborate cross-validation, some poselet candidates have high cross-validation estimate but perform poorly on real data. We evaluated the classifiers on real data and pruned 44 of them that generated the largest number of false positives

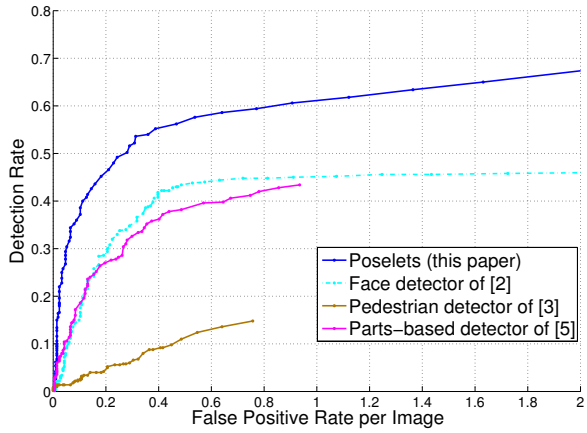

Figure 9. ROC curve comparing our torso detection performance with the frontal detector of Bourdev and Brandt [2], the pedestrian detector of Dalal and Triggs [3] and the parts-based deformable detector of Felzenszwalb, Mcallester and Ramanan [5] using the H3D test set.

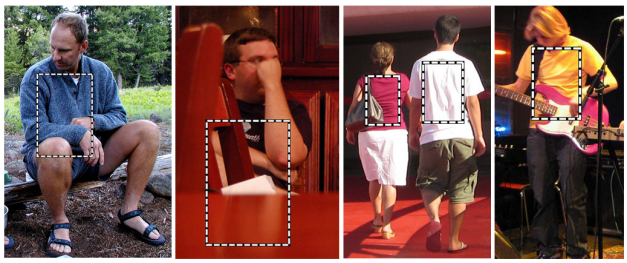

Figure 10. Examples of torso detections using poselets.

using ground truths from H3D annotations of the test set. We don't allow multiple hits to map to the same truth. Our test set is challenging; it contains partially occluded people with all kinds of poses, viewpoints and distances. Some examples of detected torsos are shown on Figure 10. We adapted the pedestrian detector of [3] and the face detector of [2] to predict torsos based on locations of found pedestrians and faces[7]. The poor performance of [3] on our test set is due not only to the difficulty of the test set but also because [3] was designed to be used for pedestrians only.

Note that pedestrian and face detectors are simply a special case of our detector when used with a single part. Our method can work with any choice of features or training methodology. In fact, our current features and classifier are identical to that of [3]; we are only limited by the size of our data set. Currently H3D generates around 2K training examples that include frontal faces (vs. 17K examples in [2]) and around 500 training examples for the pedestrian poselet (vs. 2478 in [3]).

Figure 9 also shows the performance of the part-based human detector of [5] using their implementation. Their model is trained to predict bounding boxes of people by fitting a linear regression of the bounding box from the bounding boxes of the parts. To adapt it to predict torso bounds, we used the same technique by regressing the torso bounds from the H3D training set. The figure shows that we outperform [5] on the H3D test set. To the best of our knowledge, at the time of this writing [5] has the highest score for the people category of VOC2007 at AP of 0.368 using the image context and 0.362 without context.

Detecting People on PASCAL VOC2007. We ran the poselet classifiers on VOC2007 and got AP of 0.365. For this experiment we used a simple regression of the bounds of a person from the bounds of the torso using the validation set. It is interesting that we outperform [5] on H3D but get comparable performance on VOC2007. Based on the patterns of test errors we found out that there are many cases in which our method correctly detects the person (the torso matches well) but does not predict the bounding box correctly because the person is partially occluded. In fact, if we decrease the overlap threshold from 0.50 to 0.49 our AP becomes 0.375. Our conjecture is that [5] is better at the task of detecting the *visible portions* of a person (which is what VOC2007 uses) as it requires support from the entire region. We believe our poselets classifier performs well in the presence of severe occlusion, as long as there are salient non-occluded poselets, but there is no verification stage to ensure that *all* locations agree with the torso hypothesis..

Detecting Keypoints. We trained detectors for joint and other keypoint locations using poselets. Figure 11 shows the performance. As expected, keypoints on the face have the best localization performance. The performance of shoulder keypoints is a bit worse due to symmetry of the human body. On occasion, poselets trained to detect a front facing person may detect a back-facing person and vote for the wrong shoulder, which explains why the neck is much better localized than a shoulder. Hips are much harder to localize even by humans.

7. Conclusion

We propose a two-layer classification/regression model for detecting people and localizing body components. Our first layer consists of poselet classifiers trained to detect local patterns in the image. The second layer combines the output of the classifiers in a max-margin framework. Other approaches like [5] similarly propose a two-layer model, but their part selection is unsupervised. In our work, the 3D annotation guides the search for good parts. This results in parts that are tightly clustered in configuration space and effective at joint localization.

Our method is made possible by the availability of H3D.

[7] We used their original codes, a stride of $(8, 8)$ and a scale step of 1.04 and default settings for all other parameters, except we significantly increased the margin parameters of the pedestrian detector with the hope of finding some of the closeups of people.

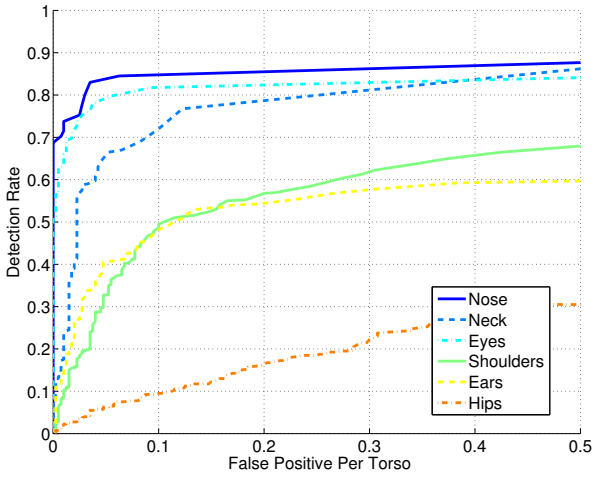

Figure 11. Detection rate of some keypoints conditioned on true positive torso detection. We consider a detection as correct if it is within $0.2S$ of its annotated location, where S is the 3D distance between the two shoulders.

Figure 12. Soft segmentation into hair, face and torso. Using H3D region annotations we have computed the pixel label probabilities of each pixel from each poselet.

The dataset can decompose the pose from the viewpoint and can generate a variety of statistics on expected pose structure, region probability masks, keypoint locations, camera view and appearance. We believe we have only scratched the surface of what H3D can do. For example, we could improve our keypoint detectors by leveraging the global pose statistics of H3D to provide configuration constraints. H3D associates pixel-level region labels with the 3D pose. We could carry the labels over to the detected poselet locations to generate a soft segmentation which we could combine with bottom-up segmentation. While we don't yet have quantitative data, an example of soft region labelling using poselets is shown on Figure 12. H3D is freely available and we hope that this will encourage researchers to exploit it in other ways.

References

[1] M. Andriluka, S. Roth, and B. Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *CVPR*, 2009.

[2] L. Bourdev and J. Brandt. Robust object detection via soft cascade. In *CVPR*, 2005.

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

[4] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge. www.pascal-network.org/challenges/VOC.

[5] P. Felzenszwalb, D. Mcallester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *CVPR*, June 2008.

[6] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *IJCV*, V61(1):55–79, January 2005.

[7] V. Ferrari, M. Marin-Jimenez, and A. Zisserman. Progressive search space reduction for human pose estimation. In *CVPR*, June 2008.

[8] B. Leibe, A. Leonardis, and B. Schiele. Combined object categorization and segmentation with an implicit shape model. In *In ECCV workshop on statistical learning in computer vision*, pages 17–32, 2004.

[9] M. Maire, P. Arbelaez, C. Fowlkes, and J. Malik. Using contours to detect and localize junctions in natural images. In *CVPR*, 2008.

[10] S. Maji and J. Malik. Object detection using a max-margin hough tranform. In *CVPR*, 2009.

[11] G. Mori and J. Malik. Estimating human body configurations using shape context matching. *ECCV*, 2002.

[12] R. Nevatia and T. O. Binford. Description and recognition of curved objects. In *Artificial Intelligence*, volume 8, pages 77–98, 1977.

[13] M. Oren, C. Papageorgiou, P. Sinha, E. Osuna, and T. Poggio. Pedestrian detection using wavelet templates. In *CVPR*, 1997.

[14] D. Ramanan. Learning to parse images of articulated bodies. *NIPS*, 2006.

[15] X. Ren, A. C. Berg, and J. Malik. Recovering human body configurations using pairwise constraints between parts. In *ICCV*, pages 824–831, 2005.

[16] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. Labelme: A database and web-based tool for image annotation. *IJCV*, 77(1-3):157–173, 2008.

[17] H. Sidenbladh and M. J. Black. Learning the statistics of people in images and video. In *IJCV*, 2003.

[18] L. Sigal and M. Black. Humaneva: Synchronized video and motion capture dataset for evaluation of articulated human motion. Technical report, Brown University, 2006.

[19] A. Sorokin and D. Forsyth. Utility data annotation with amazon mechanical turk. In *First IEEE Workshop on Internet Vision, CVPR*, pages 1–8, 2008.

[20] C. Taylor. Reconstruction of articulated objects from point correspondences in a single uncalibrated image. *Comp. Vision & Image Understanding*, 80(10):349–363, Oct 2000.

[21] B. Yao, X. Yang, and S.-C. Zhu. Introduction to a large scale general purpose ground truth dataset: methodology, annotation tool, and benchmarks. In *EMMCVPR*, 2007.