Action Recognition from a Distributed Representation of Pose and Appearance

Subhransu Maji[1], Lubomir Bourdev[1,2], and Jitendra Malik[1*]
University of California, at Berkeley[1]
Adobe Systems, Inc., San Jose, CA[2]
{smaji,lbourdev,malik}@eecs.berkeley.edu

Abstract

We present a distributed representation of pose and appearance of people called the "poselet activation vector". First we show that this representation can be used to estimate the pose of people defined by the 3D orientations of the head and torso in the challenging PASCAL VOC 2010 *person detection dataset. Our method is robust to clutter, aspect and viewpoint variation and works even when body parts like faces and limbs are occluded or hard to localize. We combine this representation with other sources of information like interaction with objects and other people in the image and use it for action recognition. We report competitive results on the PASCAL VOC* 2010 *static image action classification challenge.*

1. Introduction

We can say a fair amount about the people depicted in Figure 1 – the orientations of their heads, torsos and other body parts with respect to the camera, whether they are sitting, standing, running or riding horses, their interactions with particular objects, etc. And clearly we can do it from single image, video is helpful but not essential, and we do not need to see the whole person to make these inferences.

A classical way to approach the problem of action recognition in still images is to recover the underlying stick figure [9, 17]. This could be parameterized by the positions of various joints, or equivalently various body parts. In computer graphics this approach has been a resounding success in the form of various techniques for "motion capture". By placing appropriate markers on joints, and using multiple cameras or range sensing devices, the entire kinematic structure of the human body can be detected, localized and tracked over time [23]. But when all we have is a single image of a person, or a part of a person, not necessarily at high resolution, in a variety of clothing, the task is much harder. Research on pictorial structures [9, 17] and other techniques [19] for constructing consistent assemblies of body parts has made considerable progress, but this is very far from being a solved problem.

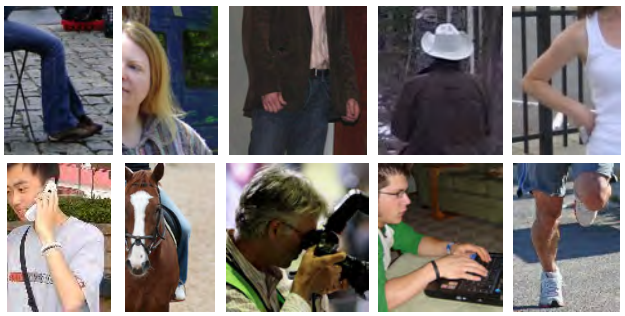


Figure 1. Pose and action is revealed from all these patches.

In this paper we take the position that recovering the precise geometric locations of various body parts is trying to solve a harder intermediate problem than necessary for our purposes. We advocate instead the use of a representation, the "poselet activation vector", which implicitly represents the configuration of the underlying stick figure, and inferences such as head and torso pose, action classification, can be made directly from the poselet activation vector.

We can motivate this by a simpler example. Consider the problem of inferring the pose of a face with respect to camera. One way of doing it is as an explicit 2D to 3D geometric problem by finding the locations of the midpoints of the eyes, nose etc, and solve for the pose. Alternatively one can consider the outputs of various face detectors - one tuned to frontal faces, another to three-quarter view faces, another to faces in profile. The responses of these detectors provide a distributed representation of the pose of the face, and one can use an "activation vector" of these responses as the input to a regression engine to estimate pose. In biological vision, strategies such as these are common place. Color is represented by a response vector corresponding to three cone types, line orientation by the responses of various simple cells in V1, and indeed neurons have been found in macaque inferotemporal cortex which show differential

*This work is supported by DOD contract W911NF-10-2-0059, Google Inc. and Adobe Systems Inc.

responses to faces at different orientations, suggesting a distributed representation there as well.

In order to generalize this strategy to the human body, we must deal with its articulated nature. Different parts can be in different configurations, and occlusion can result in only some parts being visible. In addition one needs to deal with the variation in aspect due to changes in camera direction. Poselets, introduced by Bourdev and Malik [4] and further developed in Bourdev *et al*. [3] for person detection and segmentation provide a natural framework.

We show that the poselet activation vector, which represents the degree to with each poselet is present in the image of a person, provides a distributed representation of pose and appearance. We use it to estimate the 3D orientation of the head and torso of people in the challenging PASCAL VOC 2010 person detection dataset [7]. This dataset is significantly hard where the current state of the art methods achieve detection performance only about 50%. Our approach achieves an error of 26.3° across views for the head yaw and matches the "human error rate" when the person is front facing.

Action recognition from still images can benefit from this representation as well. Motion and other temporal cues which have been used for generic action recognition from videos [20, 22, 11, 6], are missing in still images which makes it a difficult problem. In this setting the pose and appearance of the person provides valuable cues for inferring the action. For example as seen in Figure 2, certain actions like walking and running are associated with specific poses while people riding bikes and horses have both a distinctive pose and appearance.

Actions often involve interactions with other objects and one can model these interactions to disambiguate actions [26]. In addition context based on actions of other agents in the scene can provide valuable cues as well [13]. For example, certain activities like marathon events or musicians playing in a concert, are group activities and it is likely that everyone in the scene is performing the same action.

The rest of the paper is structured as follows: we begin with a review of work in the area of action recognition and pose estimation in Section 2. In Section 3, we describe how we construct the poselet activation vector for a given person in an image. We present experiments on 3D pose estimation of people in the PASCAL VOC 2010 people detection dataset in Section 4. Finally we report results on the recently introduced PASCAL VOC 2010 action classification dataset in Section 5 and conclude in Section 6.

2. Previous Work

The current work draws from the literature of two active areas in the computer vision – pose estimation and action recognition. We briefly describe some without any hope of doing justice to either of the areas.

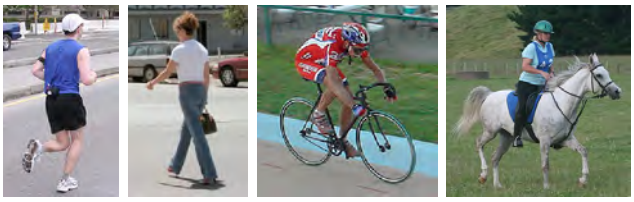


Figure 2. Pose and appearance variation across actions.

Human pose estimation from still images. Pictorial structures based algorithms like that of [9, 19, 17, 10] deal with the articulated nature of humans by finding body parts like limbs and torsos and constructing the overall pose using the prior knowledge of human body structure. Though completely general, these methods suffer when the parts are hard to detect in images. Another class of methods work by assuming that the humans appear in backgrounds which are easy to remove, and in such cases the contour carries enough information about the pose. This includes the shape-context based matching of silhouettes in the work of [16], the work of [21] where approximate nearest neighbor techniques are used to estimate the pose using a large dataset of annotated images.

A common drawback of all these approaches is that they treat the task of pose estimation and detection separately. Pictorial structure based models often assume a rough localization of the person and fail when there is significant occlusion or clutter. In such a two-stage pipeline it would be helpful if the detector provides a rough estimate of the pose to guide the next step. We also believe that the detection algorithms need to have a crude treatment of pose in them. This is reflected by the fact that some of the best people detectors on the PASCAL VOC challenge namely the detector of Felzenszwalb *et al*. [8] and Bourdev *et al*. [3] are part based detectors which have some treatment of pose.

Action Recognition from video. Actions in this setting are described by some representation of its spatio-temporal signature. This includes the work of Blank *et al*. [2] and Shechtman and Irani [22], who model actions as space-time volumes and classification is based on similarity of these volumes. Schuldt *et al*. [20] and Laptev [14] generalize the notion of interest points from images to space-time volumes and use it to represent actions. Actions as motion templates has been explored in the work of Efros *et al*. [6], where actions are described as series of templates of optical flow. Other methods like [18, 27] are based on representations on the 2D motion tracks of a set of features over time.

Action recognition from still images. Humans have a remarkable ability to infer actions from a still image as shown in Figure 1. In this setting it is natural to build representations on top the output of a pose estimation algorithm. Due to the drawbacks of the current pose estimation algorithms,

Figure 3. **Our distributed representation of pose using poselets.** Each image is shown with the top 9 active poselets consistent with the person in the image (shown by their average training examples). Occlusion, variations in clothing, clutter, lack of resolution in images makes the pose estimation a hard problem and our representation is robust to these.

several approaches build pose representations that are more robust – Ikizler and Pinar [12] represent pose using a "histogram of oriented rectangle" feature which is the probability distribution of the part locations and orientations estimated using part detectors. Thurau and Hlavac [24] represent pose as a histogram of pose primitives. These methods inherit most if not all of the problems of pose estimation.

The closest in spirit to our approach is the work of Yang *et al*. [25], who also use a representation based on poselets to infer actions. Pose represented as a configuration of body part locations is expressed as a latent variable which is used for action recognition. Training and inference in the model amount to reasoning over these latent poses which are themselves inferred using a tree like prior over body parts and poselet detections. Unlike their approach we don't have an explicit representation of the pose and use the "poselet activation vector" itself as a distributed representation. In addition, our poselets encode information from multiple scales and are not restricted to parts like legs and arms. In our experiments we found that such an overcomplete representation greatly improves the robustness of the system. We show that linear classifiers trained on top of the poselet activation vector can be used for both 3D pose estimation of people in the challenging PASCAL VOC 2010 dataset and static image action recognition demonstrating the effectiveness of our representation.

3. Poselet Activation Vector

Our framework is built on top of poselets [4, 3] which are body part detectors trained from annotated data of joint locations of people in images. The annotations are used to find patches similar in pose space to a given configuration of joints. A poselet is a SVM classifier trained to recognize such patches. Along with the appearance model one can also obtain the distributions of these joints and person bounding boxes conditioned on each poselet from the annotations. Figure 10 shows some example poselets.

Given the bounding box of a person in an image, our representation, called the poselet activation vector, consists of poselets that are consistent with the bounding box. The vector has an entry for each poselet type which reflects the degree to which the poselet type is active in that person. This provides a distributed representation of the high dimensional non-linear pose space of humans as shown in Figure 3. Notice that the pose and appearance information is encoded at multiple scales. For example, we could have a part which indicates just the head or just the torso or the full pedestrian. We use this representation for both action recognition and 3D pose estimation from still images.

4. 3D Pose Estimation from Still Images

First we quantitatively evaluate the power of the poselet activation vector representation for estimating pose. Our task is to estimate the 3D pose of the head and torso given the bounding box of the person in the image. Current approaches for pose estimation based on variants of pictorial structures are quite ill suited for this task as they do not distinguish between a front facing and back facing person. Some techniques can estimate the 3D pose of the head by first detecting fiducial points and fitting it to a 3D model of the head, or by regressing the pose from the responses of face detectors trained to detect faces at different orientations [15]. These methods are not applicable when the face itself is occluded or when the image is at too low a resolution for a face detector, a common occurrence in our dataset.

The pose/aspect of the person in encoded at multiple scales and often one can roughly guess the 3D pose of the person from various parts of the person as seen in Figure 1 and our representation based on poselets are an effective way to use this information. Our results show that we are able to estimate the pose quite well for both profile and back facing persons.

A Dataset of 3D Pose Annotations. Since we wanted to study the problem of pose estimation in a challenging setting, we collected images of people from the *validation* subset of PASCAL VOC 2010 dataset not marked as difficult. We asked the users on Amazon Mechanical Turk [1], to estimate the rotations around X,Y and Z of the head and torso by adjusting the pose of two gauge figures as seen in Figure 4(a). We manually verified the results and threw away the images where there was high disagreement between the

annotators. These typically turned out to be images of low resolution or severe occlusion.

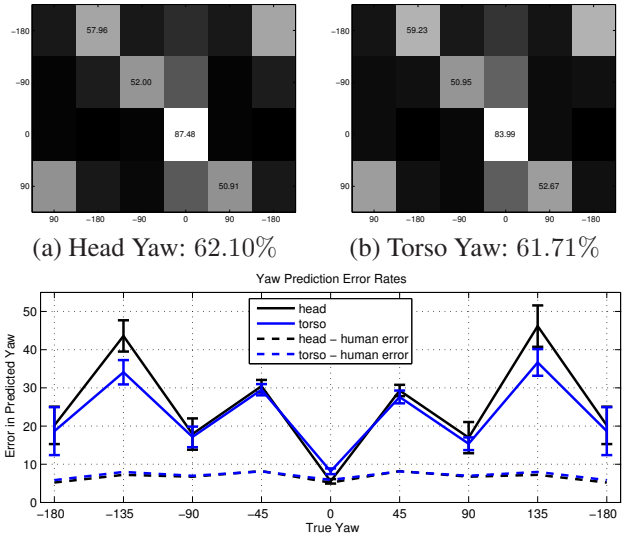
Our dataset has very few examples where the rotation along X and Z axes is high, as is typical of consumer photographs, hence we removed images which have rotations along X and Z $> 30^\circ$ and focus on estimating the rotation around Y (Yaw) only. In the end we have 1620 people annotations that along with their reflections result in 3240 examples. The distribution of the yaw across the dataset is shown in Figure 4(c, d, e).

Figure 4(b) shows the human error in estimating the yaw across views of the head and torso. This is measured as the average of standard deviation of the annotations on a single image in the view range. The error is small for people in canonical views, i.e. when the person is facing front, back, left or right, whereas it is high when the person in facing somewhere in between. Overall the annotators are fairly consistent with one another with a median error of 6.66° for the head and 7.07° for the torso across views.

Experiments. Similar to [3], we train 1200 poselets on the PASCAL train 2010 + H3D trainval dataset. Instead of all poselets having the same aspect ratio, we used four aspect ratios: 96×64, 64×64, 64×96 and 128×64 and trained 300 poselets of each. In addition we fit a model of bounding box prediction for each poselet. We construct the poselet activation vector by considering all poselet detections whose predicted bounding box overlaps the bounding box of the person, defined by the intersection over union > 0.20 and adding up the detection scores for each poselet type. We use this 1200 dimensional vector to estimate the pose of the person.

We estimate the pose of the head and torso separately. We discretize the yaw $\in [-180^\circ, 180^\circ]$ into 8 discrete bins and train one-vs-all linear classifiers for predicting the discrete label. The angle is obtained by parabolic interpolation using the highest predicted bin and its two adjacent neighbors. We optimize our parameters on one split of the data and report results using 10 fold cross validation. We split the training and test set equally ensuring both the image and its reflection are both either in the training or the test set.

Figure 5(a, b) show the confusion matrix for the task of predicting the discrete view, one of front, left, right and back, for the head and torso. The average diagonal accuracy is 62.1% for the head and 61.71% for the torso. The median errors in predicting the real valued view are shown in Figure 5(c). We report results by averaging the error for predicting view across 8 discrete views. Since the dataset is biased towards frontal views, this error metric gives us a better idea of the accuracy of the method. Across all views the error is about 26.3° and 23.4° for the head and torso respectively, while across the front views, i.e. yaw $\in [-90^\circ, 90^\circ]$, the error is lower: 20.0° and 19.6° respectively. In particular, the error when the person is facing front, i.e. yaw $\in [-22.5^\circ, 22.5^\circ]$ matches the human error rate. Our method is able to recognize the pose of back facing people, i.e. yaw $\in [157.5^\circ, -157.5^\circ]$, a 45° range around the back facing view, with an error of about 20° error for the head and torso. Approaches based on face detection would fail but our representation benefits from information at multiple scales like the overall shape of the person, as shown in Figure 6. The error is smaller when the person is facing exactly left, right, front and back while it is higher when the person is facing somewhere in between, qualitatively similar to humans.



(a) Head Yaw: 62.10% (b) Torso Yaw: 61.71%

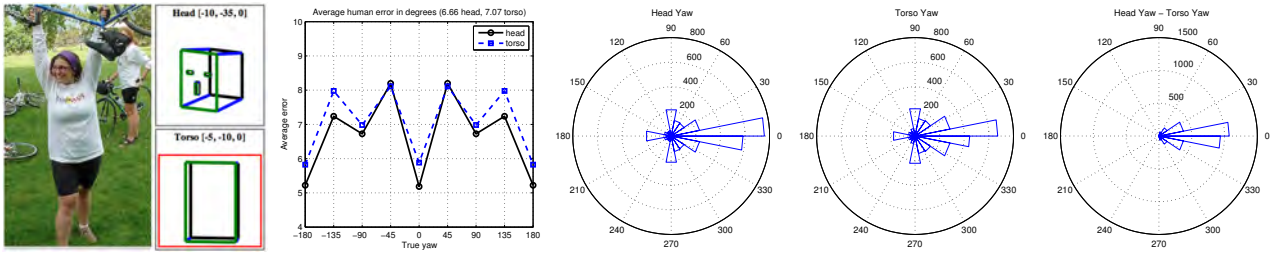
(c) Error in predicting yaw across views

Figure 5. (a, b) Average confusion matrix over 10-fold cross validation, for predicting four views *left, right, front* and *back*. The mean diagonal accuracy is 62.10% and 61.71% for predicting the head and the torso respectively. (c) Error in predicting the yaw averaged over 8 discrete views using 10-fold cross validation. Across all views the error is about 26.3° and 23.4° for the head and torso respectively, while across the front views, i.e. yaw $\in [-90^\circ, 90^\circ]$, the error is lower $20.0^\circ, 19.6^\circ$. In particular the error when the person is facing front, i.e. yaw $\in [-22.5^\circ, 22.5^\circ]$ matches the human error rate.

At roughly 25° error across views, our method is significantly better than the baseline error of 90° for the method that always predicts the view as frontal (It gets 0° error for frontal view, but 180° error for back view). Figure 7 shows some example images in our dataset with the estimated pose. We believe this is a good result on this difficult dataset demonstrating the effectiveness of our representation for coarse 3D pose estimation.

5. Static Action Classification

In this section we present our method for action classification and report results on the newly introduced PASCAL VOC 2010 action classification benchmark. The in-



(a) AMT interface (b) Human Error (c) Head Yaw (d) Torso Yaw (e) Head-Torso Yaw

Figure 4. (a) Interface for annotating the 3D pose on Amazon Mechanical Turk. (b) Human error rate across view for estimating the pose of the head and torso. (c, d, e) Distribution of the yaw of head, torso and torso relative to the head, on our 3D pose dataset.

Figure 7. Left to right are examples images in our 3D pose dataset of increasing prediction error. Under each image the plot shows the true yaw for the head (left) and torso (right) in green and the predicted yaw in red. We are able to estimate the pose even when the face, limbs and other body parts are hard to detect.

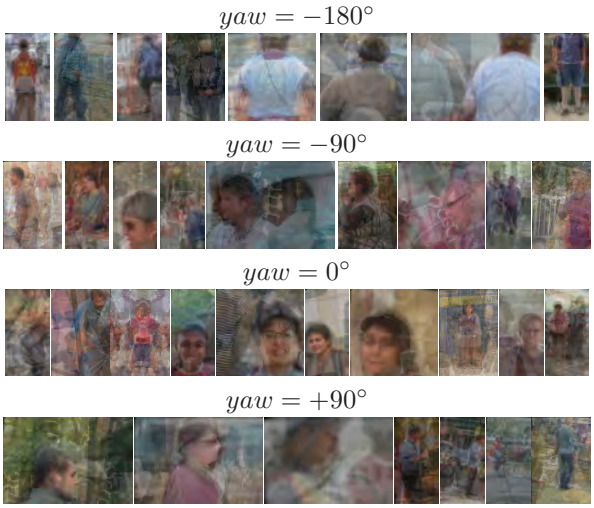


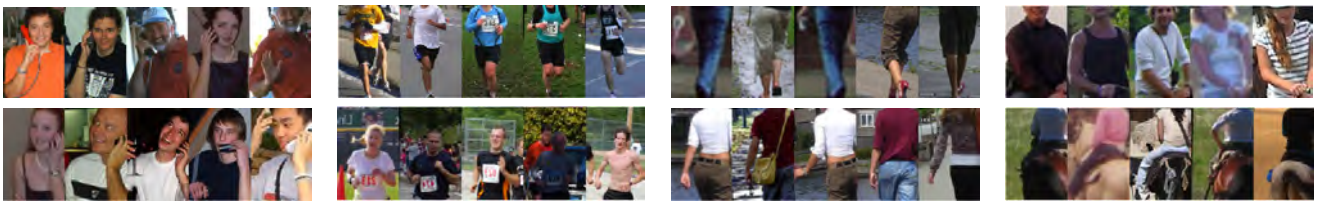
Figure 6. **Poselets with the highest weights for discrete view classification of the head.** Note that information from multiple scales is used to infer the view. When the person is back-facing, i.e. yaw $= -180°$, poselets corresponding to pedestrians and upper-body are selected where as for the frontal view face poselets are selected.

put is a set of bounding boxes on images and the task is to score each of these with respect to nine action categories namely : *phoning, playinginstrument, reading, ridingbike, ridinghorse, running, takingphoto, usingcomputer* and *walking*. Figure 2 shows examples from various action categories.

Action specific poselets. There are 608 training examples for all the action categories. To train poselet models we first annotate each person with 2D joint locations on Amazon Mechanical Turk. Five independent annotators were asked to annotate every image and the results were averaged with some outlier rejection. Similar to the approach of [3] we randomly sample windows of various aspect ratios and use the joint locations to find training examples each poselet.

Figure 8 shows that pose alone cannot distinguish between actions and the appearance information is complimentary. For example we would like to learn that people riding bikes and horses often wear helmets, runners often wear shorts, or that people taking pictures have their faces occluded by a camera. *To model this, we learn action specific appearance by restricting the training examples of a poselet to belong to the same action category.*

Many poselets like a "face" poselet may not discriminate between actions. *The idea illustrated in Figure 9, is windows that capture salient pose specific to certain actions are likely to be useful for action discrimination.* We measure "discriminativeness" by the number of within class examples of the "seed" windows in the top $k = 50$ nearest examples for the poselet. The idea is that if a pose is discriminative then there will be many examples of that poselet from within the same class. Combined with the earlier step this gives us a way to select poselets which detect salient pose and appearance for actions as shown in Algorithm 1. Appearance models are based on HOG [5] and linear SVM. We learn 1200 action specific poselets. Figure 10 shows representative poselets from four action categories.



phoning *running* *walking* *ridinghorse*

Figure 10. Example poselets shown by their top 5 training examples for various action categories. These capture both the pose and appearance variation across the action categories.

Algorithm 1 Action specific poselet selection.

Require: 2D keypoint/action labels on training images.

1: **for** $i = 1$ to n **do**
2: Pick a random seed window and find the nearest examples in configuration space based on the algorithm of [3].
3: Compute the number of within class examples in the $k = 50$ nearest examples.
4: **end for**
5: Select the top m seed windows which have the highest number within class examples.
6: For each selected window, restrict the training examples to within the class and learn an appearance model based on HOG and linear SVM.

Remarks:

- *Steps* $1-5$ *learn action specific pose, while step* 6 *learns action specific appearance.*
- *We ensure diversity by running steps* $1-6$ *in parallel. We set* $m = 60, n = 600$ *across* 20 *nodes to learn* 1200 *poselets.*

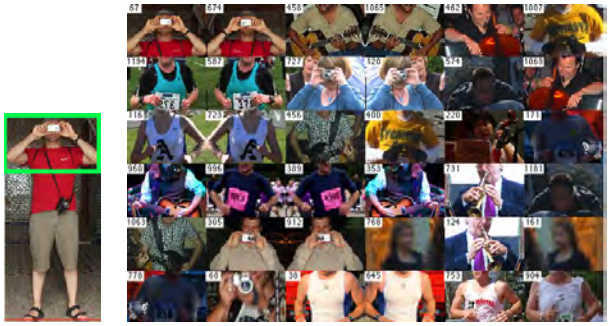
seed all examples within class examples

Figure 8. The middle image shows the nearest examples matching the seed using the pose alone, while the image on right shows the top examples within the *takingphoto* category. This allows us to learn appearance and pose specific to that action.

Poselet Activation Vector. The action poselets are run in a scanning window manner and we collect poselet detections whose predicted bounds overlap the given person bounds, defined by the intersection over union of the area > 0.15. The i'th entry of the poselet activation vector is the sum of the scores of all such detections of poselet type i.

Spatial Model of Object Interaction. Interaction with other objects often provides useful cues for disambiguating actions [26]. For example, images of people riding horses have the person and the horse in certain spatial configurations. We model the interaction with four object categories : *horse, motorbike, bicycle* and *tvmonitor*. We learn a mixture model of the relative spatial location between the person bounding box and the object bounding box in the image as shown in Figure 11. For detecting these objects we use the detector based on poselets trained on these object categories presented in the PASCAL VOC 2010 object detection challenge. For each object type we fit a two component mixture model of the predicted bounding box to model the various aspects of the person and objects.

Given the object detections we find all the objects whose predicted person bounds overlap the bounds of the given



seed Top 36 training examples

seed Top 36 training examples

Figure 9. The top row shows a seed window that captures a salient pose for the *takingphoto* category. The 36 nearest examples in configuration space for the top seed window has 7 examples from the *takingphoto* category while the bottom seed has only 2.

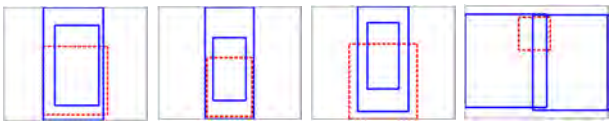


Figure 11. Spatial model of the object person interaction. Each image shows the modes of the bounding boxes of the person (blue) relative to the bounding box of the object (red). For *horse, motorbike* and *bicycle* category the two modes capture front and side views of the object while for the *tvmonitor* it captures the fact that TV monitors are often at the left or right corner of the person bounding box.

person > 0.3. Similar to the poselet activation vector we construct an "object activation vector" by taking the highest score of the detection for each object type among these.

Action context. Often the action of a person can be inferred based on what others are doing in the image. This is particularly true for actions like *playinginstrument* and *running* which are group activities. Our action context for each person is a 9 dimensional vector with an entry for each action type whose value is the highest score of the action prediction among all the other people in the image. Overall the second stage classifier is a separate linear SVM for each action type trained on 10 features: self score for that action and 9 for action context.

Experiments. Table 1 shows the performance of various features on the test and validation set. All the parameters described were set using a 10-fold cross validation on the trainval subset of the images.

The *poselet activation vector* alone achieves a performance of 59.8 on the validation subset of images and does quite well in distinguishing classes like *ridinghorse, running, walking* and *phoning*. Adding the object model boosts the performance of categories like *ridingbike* and *usingcomputer* significantly, improving the average AP to 65.3. These classes either have the widely varying object types and poselets are unable to capture the appearance variation. Modeling the spatial interaction explicitly also helps for classifying *usingcomputer* class as the interaction is often outside the bounding box of the person. Finally the context based re scoring improves the performance of *playinginstrument* and *running* class as these are often group activities.

Figure 12 shows the confusion matrix of our classifier. Some high confusion pairs are {*reading, takingphoto*} $\rightarrow$ *playinginstrument* and *running* $\rightarrow$ *walking*. Figure 13 shows misclassified examples for several pairs of categories. Overall our method achieves an AP of 65.6 on the validation and 59.7 on the test set which is comparable to the winning techniques in PASCAL VOC 2010 challenge, for example, 60.1 for "INRIA_SPM_HT" and 60.3 for "CVC_BASE". We refer the readers to the challenge website[1] for details and results of other entries.

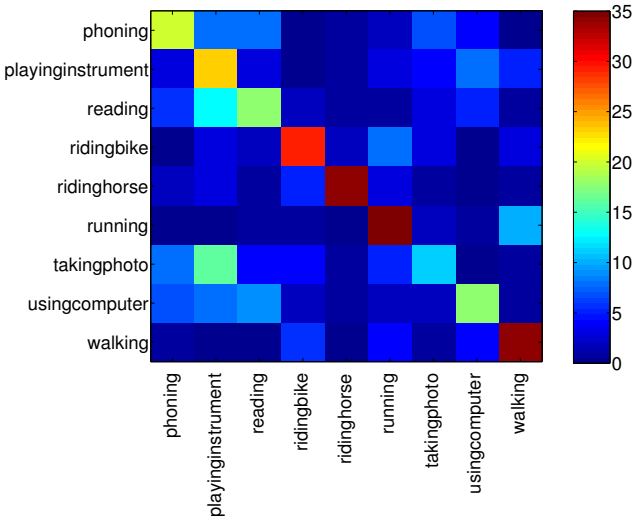


Figure 12. Confusion matrix for our action classifier. Each row shows the distribution of the true labels of the top 50 ranked examples for each action category on the validation subset of the images. Some high confusion pairs are {*reading, takingphoto*} $\rightarrow$ *playinginstrument* and *running* $\rightarrow$ *walking*.

Figure 13. **Pairwise confusions between several classes on the PASCAL 2010 validation set**. Each A $\rightarrow$ B shows the top 4 images of class A ranked by classifier of class B. Confusion is often caused when the person has similar pose or failures of the object detector.

[1] http://pascallin.ecs.soton.ac.uk/challenges/VOC

	Validation			Test
category	PAV	w/ OAV	w/ C	w/ C
phoning	63.3	62.0	62.0	49.6
playinginstrument	44.2	44.4	45.6	43.2
reading	37.4	44.4	44.3	27.7
ridingbike	62.0	84.7	85.5	83.7
ridinghorse	91.1	97.7	97.5	89.4
running	82.4	84.1	86.0	85.6
takingphoto	21.1	22.9	24.6	31.0
usingcomputer	54.2	64.9	64.3	59.1
walking	82.0	83.6	80.8	67.9
average	59.8	65.3	65.6	59.7

Table 1. Average precision on the action validation and test set using various features. PAV is the performance using just the *poselet activation vector*. Column w/OAV shows the performance by including the *object activation vector* as features and column w/C shows the performance by including action context. The object features help in the *ridingbike, ridinghorse* and *usingcomputer* categories, while the context improves the performance on *playinginstrument* and *running* categories. Our methods achieves an average AP of 59.7 on the test set which is comparable to the winning techniques in PASCAL VOC 2010.

6. Conclusion

We demonstrate the effectiveness of the poselet activation vector on the challenging tasks of 3D pose estimation of people and static action recognition. Contrary to the traditional way of representing pose which is based on the locations of joints in images, we use the poselet activation vector to capture the inherent ambiguity of the pose and aspect in a multi-scale manner. This is well suited for estimating the 3D pose of persons as well as actions from static images. In the future we would like to investigate this representation for localizing body parts by combining top down pose estimates with bottom-up priors and exploit pose-to-pose constraints between people and objects to estimate pose better.

Most of the other high performing methods on the PASCAL VOC 2010 action classification task use low-level features based on color and texture together with a SVM classifier, without any explicit treatment of pose. We believe that such methods benefit from the fact that one is provided with accurate bounding boxes of the person in the image. This is quite unrealistic in an automatic system where one has to estimate the bounds using a noisy object detector. We on the other hand use the bounding box information quite loosely by considering all poselet detections that overlap sufficiently with the bounding box. In addition, the poselet activation vector provides a compact representation of the pose and action, unlike the high dimensional features typical of "bag-of-words" style approaches.

The annotations and code for estimating the yaw of the head and torso in images, as well as the keypoint annotations and code for static image action classification can be downloaded at the author's website.

References

[1] Amazon mechanical turk. http://www.mturk.com. 3179

[2] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In *ICCV*, 2005. 3178

[3] L. Bourdev, S. Maji, T. Brox, and J. Malik. Detecting people using mutually consistent poselet activations. In *ECCV*, sep 2010. 3178, 3179, 3180, 3181, 3182

[4] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *ICCV*, 2009. 3178, 3179

[5] N. Dalal and B. Triggs. Histograms of Oriented Gradients for Human Detection. In *CVPR*, 2005. 3181

[6] A. A. Efros, A. C. Berg, G. Mori, and J. Malik. Recognizing action at a distance. *ICCV*, 2, 2003. 3178

[7] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2):303–338, June 2010. 3178

[8] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *TPAMI*, 32(9):1627 –1645, 2010. 3178

[9] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *Int. J. Comput. Vision*, 61:55–79, January 2005. 3177, 3178

[10] V. Ferrari, M. Marin-Jimenez, and A. Zisserman. Progressive search space reduction for human pose estimation. In *CVPR*, 2008. 3178

[11] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. *TPAMI*, 29, Dec. 2007. 3178

[12] N. Ikizler and P. Duygulu. Histogram of oriented rectangles: A new pose descriptor for human action recognition. *Image Vision Comput.*, 27, September 2009. 3179

[13] T. Lan, Y. Wang, W. Yang, and G. Mori. Beyond actions: Discriminative models for contextual group activities. In *NIPS*, 2010. 3178

[14] I. Laptev. On space-time interest points. *Int. J. Comput. Vision*, 64:107–123, September 2005. 3178

[15] K. Mikolajczyk, R. Choudhury, and C. Schmid. Face detection in a video sequence-a temporal approach. In *CVPR*, volume 2, pages II–96. IEEE, 2001. 3179

[16] G. Mori and J. Malik. Recovering 3d human body configurations using shape contexts. *TPAMI*, 28(7), 2006. 3178

[17] D. Ramanan and D. Forsyth. Finding and tracking people from the bottom up. In *CVPR*, volume 2, 2003. 3177, 3178

[18] C. Rao, A. Yilmaz, and M. Shah. View-invariant representation and recognition of actions. *Int. J. Comput. Vision*, 50, November 2002. 3178

[19] X. Ren, A. Berg, and J. Malik. Recovering human body configurations using pairwise constraints between parts. In *ICCV*, volume 1, 2005. 3177, 3178

[20] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: a local svm approach. In *ICPR*, volume 3, 2004. 3178

[21] G. Shakhnarovich, P. Viola, and T. Darrell. Fast pose estimation with parameter-sensitive hashing. In *CVPR*, 2003. 3178

[22] E. Shechtman and M. Irani. Space-time behavior based correlation or how to tell if two underlying motion fields are similar without computing them? In *TPAMI*, volume 29, pages 2045–2056, November 2007. 3178

[23] L. Sigal, A. Balan, and M. Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated humanmotion. *IJCV*, 87:4–27, 2010. 10.1007/s11263-009-0273-6. 3177

[24] C. Thurau and V. Hlavac. Pose primitive based human action recognition in videos or still images. In *CVPR*, 2008. 3179

[25] W. Yang, Y. Wang, and G. Mori. Recognizing human actions from still images with latent poses. In *CVPR*, 2010. 3179

[26] B. Yao and L. Fei-Fei. Modeling mutual context of object and human pose in human-object interaction activities. In *CVPR*, San Francisco, CA, June 2010. 3178, 3182

[27] A. Yilmaz and M. Shah. Actions sketch: a novel action representation. In *CVPR*, volume 1, 2005. 3178