Robust Pose Features for Action Recognition

Hyungtae Lee, Vlad I. Morariu, Larry S. Davis
University of Maryland, College Park MD USA
htlee,morariu,lsd@umiacs.umd.edu

Abstract

We propose the use of a robust pose feature based on part based human detectors (Poselets) for the task of action recognition in relatively unconstrained videos, i.e., collected from the web. This feature, based on the original poselets activation vector, coarsely models pose and its transitions over time. Our main contributions are that we improve the original feature's compactness and discriminability by greedy set cover over subsets of joint configurations, and incorporate it into a unified video-based action recognition framework. Experiments shows that the pose feature alone is extremely informative, yielding performance that matches most state-of-the-art approaches but only using our proposed improvements to its compactness and discriminability. By combining our pose feature with motion and shape, we outperform state-of-the-art approaches on two public datasets.

1. Introduction

Action recognition still remains challenging due to great intra and inter variance of classes, cluttered and occluded background, etc., despite numerous recent advances. Many researchers extract local image and video features from video sequences, separate them into clusters, and generate histogram-based representations. Interest points are often extracted by methods such as Harris3D [9], Hessian [19], etc, to capture shape and motion of local points. HOG [8], silhouettes [8], and SIFT [12] are commonly used as shape features. As a motion feature, most researchers use optical flow [8] or other custom representations of space-time volumes, e.g., Liu et al. [12] use flat gradients within 3D cuboids.

While pose-based action recognition methods have also been studied [7], they have generally underperformed methods based on shape and motion features on difficult "in-the-wild" videos such as those obtained from YouTube. This is because pose estimation remains a difficult problem in uncontrolled settings and even state-of-the-art pose estimation approaches are relatively brittle.

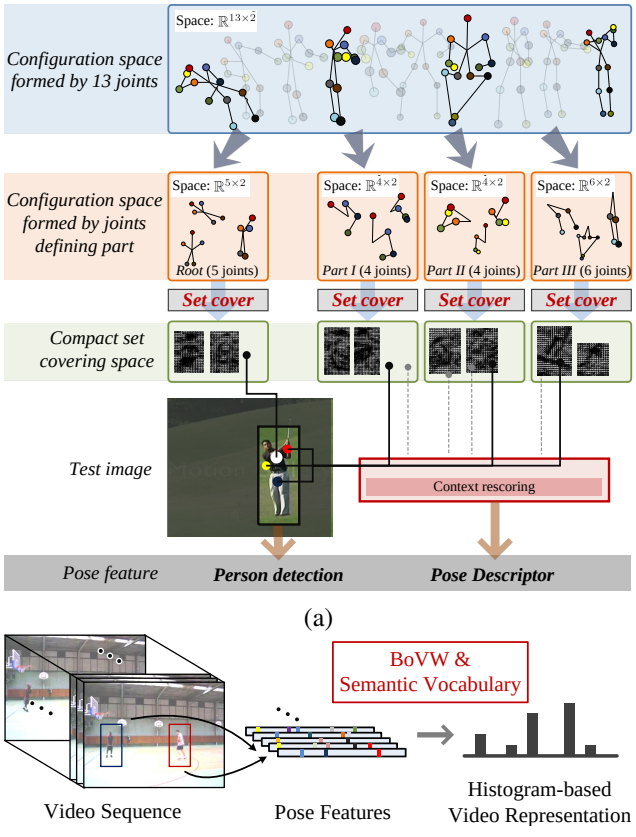

Figure 1. Illustration of our proposed posed descriptor and its use for action recognition. In (a) the 13 joint pose configuration space is split into subsets of joints, whose smaller space of configurations we cover greedily with poselet models. This ensures that common *and* rare configurations are represented (*covered*). This improves action recognition which models transitions through pose configurations. Given an image, poselet activations are obtained, as usual, grouped by mutual consistency, and assembled into an activation vector, which is rescored to incorporate the context provided by mutually consistent activations. In (b) we depict the use of the proposed descriptor in a histogram based video representation.

In this work, we use a pose feature based on poselets, which captures human pose without the need for exact lo-

calization of joint locations, but instead relies on the representation and detection of coarse qualitative poses (e.g., standing, bending) which are learned automatically from training data. Poselets [2, 4] are discriminative part models constructed to be tightly clustered in the configuration space of joints as well as in the appearance space of images, and which have been successfully used for detecting people [2, 4], describing human attributes [3], and recognizing human actions [13, 21, 22] in single images. As more poselets are used by an object detector, the detector's accuracy increases, but its efficiency decreases proportionally with the number of poselets.

While a small number of poselets might be sufficient for detection, for action recognition it becomes important to cover the space of pose variations more completely, since actions are generally modeled as transitions through pose space. However, the standard poselets training procedure requires too many poselets to adequately represent the pose space for action recognition. This leads to a loss in efficiency, increases the feature descriptor size, and ultimately leads to poor action recognition performance (as shown in our experiments). This motivates us to modify the poselet training procedure with the following goals in mind: (1) increase the coverage of the space of poses, and (2) maintain efficiency by making the set of poselets more compact. To accomplish this we partition the 13 joints into overlapping subsets (depicted in Figure 1), and instead of randomly selecting image rectangles to define poselets as in [4], we select seed rectangles using greedy set-cover to ensure that most joint configurations in each subset are adequately detected by a poselet. Our proposed greedy set cover algorithm ensures that each part–defined as a subset of joints–should generate poselets that cover the entire range of its configurations while avoiding redundant poselets (each poselet should detect at least one new configuration that is not detected by another poselet).

Given a test video, we obtain a pose descriptor from our compact set of poselets by constructing activation vectors from mutually consistent activations as in [4], and rescore activations using the context encoded by this vector. We construct activation vectors for each root activation and create a codebook based histogram representation using all root activations that have a high enough confidence after context rescoring. We incorporate the proposed pose features in existing action recognition [12] with traditional motion and shape features.

Figure 1 depicts our approach. To summarize, our contributions are the following: 1) we improve the compactness and discriminability of the original poselets by a training process that applies greedy set cover to the smaller configuration spaces of joint subsets, and 2) we are the first to our knowledge to successfully use pose as a feature for "in-the-wild" video-based action recognition.

We evaluate our approach on two benchmarks: YouTube sports dataset [15] and YouTube action dataset [12]. Our experiments show that the proposed pose feature provides significant complementary information to the motion and shape features. In fact, the pose feature alone nearly matches state-of-the-art results, while the combination with either shape or motion alone improves over the state-of-the-art, and the combination of all three types of feature outperforms all other alternatives. In fact, on the YouTube Action dataset, our proposed approach outperforms the state-of-the-art by over 10%. Our experiments demonstrate the importance of our modified training procedure to effectively incorporate poselet features into a video-based action recognition framework.

In section 2, we discuss related work. In section 3 and 4, we describe details of semantic pose features and incorporating features into an action recognition framework, respectively. In section 5, we present the experimental results that demonstrate the performance of our approach. We present our concluding remarks in section 6.

2. Related Work

Since the literature on action recognition is vast, we describe only recent works in this section. Liu et al. [12] extract motion and shape features from videos, construct a compact yet discriminative visual vocabulary using an information-theoretic algorithm, and generate a histogram-based video representation. While this approach is effective, it does not make use of pose features. We extend this approach by incorporating our proposed pose feature to their features and followed the framework for action recognition proposed by [12]. Xie et al. [20] explore the use of deformable part models (DPM) for incorporating human detection and pose estimation into action recognition. Similar to our method, their work is also based on human poses but our part models are trained to discriminate between various poses of a person, unlike DPM's, which are trained to discriminate between patches in which a person is present or absent. Le et al. [10] learn features directly from video using independent subspace analysis that is robust to translation and selective to frequency and rotation changes. Todorovic [17] views a human activity as a space-time repetition of activity primitives and models the primitives and their repetition by a generative model-graph. Sadanand and Corso [16] propose action bank, consisting of action detectors sampled according to classes and viewpoints.

Our proposed pose feature is based on the poselets framework introduced by Bourdev and Malik [4]. Poselets are discriminative part detectors constructed from tight clusters in the configuration space of the human articulated body as well as in the appearance space of images. At test time, poselet activations are detected by multi-scale sliding windows, and persons are detected by Max Margin

Hough Voting [4] or by clustering mutually consistent activations [2]. Poselets have been employed to improve results in various vision applications, including segmentations [2], subordinate categorization [6], attribute classification [3], pose [11, 13] and action recognition [13, 21, 22]. Unlike all of these extensions of poselets which are applied to static images, our method extends the use of poselets to action recognition on video sequences, producing results that improve on the current state-of-the-art.

3. Training Parts and Context Rescoring

3.1. Motivation

Poselets are successfully used in detecting humans [4] as well as recognizing actions [13] in still images but have not been used for video-based action recognition. While a small number of poselets might be sufficient for detection, for action recognition it becomes important to cover the space of pose variations more completely, so that we can observe and model transitions through the pose space. However, if the number of poselets is increased, person detection by clustering consistent activations may be impractical since the clustering complexity is quadratic in the number of poselet activations.

We modify the poselet training procedure in three ways to improve its effectiveness and efficiency. First, we manually select three sets of joints predictive of pose and introduce three parts that cover the extents of those joints in each set. We also select a set of joints corresponding to the head and torso that are stable and are suitable for use as a root for our model (similar to the root in DPM models [7], which serves as a coarse description of the person). Second, we modify the procedure for selecting a poselet seed, replacing random selection with greedy set cover to satisfy the following criteria:

1. **effectiveness:** each part should generate poselets that cover the entire range of its potential configurations,
2. **efficiency:** poselets should not be redundant.

Third, instead of clustering pairs of mutually consistent poselets to obtain detections of people, we use all root activations as potential human detections, and rescore them out by training a classifier on the feature vector containing the activation scores of the root candidate and of the parts consistent with that root candidate. This yields a clustering process whose computational requirements increase linearly (instead of quadratically) with the number of part activations, allowing for the use of a larger number of poselets in our framework.

3.2. Definition of Parts and Training Poselets

Definition of parts: We follow the definition of the root and the parts in [21] employing a four part star structured

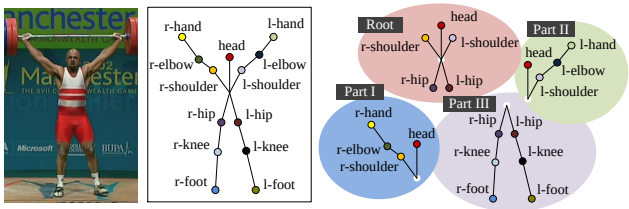

Figure 2. Joints annotation (left) and definition of root and parts (right).

Combination of joints	Proc. dist	Coverage
l_shoulder-l_elbow-l_hip-l_knee	0.6178	0.5255
l_shoulder-l_elbow-l_hand-l_hip	1.1526	0.5658
head-l_shoulder-l_hip-l_knee	0.4509	0.5461
head-l_shoulder-l_elbow-l_hip	0.5980	0.6266
head-l_shoulder-l_elbow-l_hip	0.2490	0.6637
head-l_shoulder-l_elbow-l_hip-l_knee	0.4789	0.5238
head-l_shoulder-l_elbow-l_knee-l_hip	0.7819	0.5641
head-l_shoulder-r_shoulder-l_hip-r_hip	**0.1390**	**0.6566**

Table 1. Combinations of joints which appear in more than 50 % of YouTube sports dataset [15] are selected and procrustes distance among configurations of each combination are computed. The joints that define our root (in bold) achieve the best trade-off between joint location stability and dataset coverage.

model to express human pose for recognizing actions. The root is defined by the head, shoulders, and hips and the three parts are defined by pairs of limbs: (head, right shoulder, right elbow, right hand), (head, left shoulder, left elbow, left hand), and (hips, knees, feet) (Fig. 2). Table 1 shows the average procrustes distance among pairs of training configurations, as well as the coverage of poselets trained on these joints. The table provides the experimental support for using the combination of the head, shoulders, and hips as a root. Only the activation vector of the root is rescored and used in the descriptor, since its coverage is high while the joints belonging to the root are relatively stable, as shown by the low procrustes distance among the root joints.

Training poselets: The appearance variations of the root and each part are captured by multiple poselets trained by covering the configuration space of each part. Each poselet is trained by the process described in [2]. The patch (seed of a poselet) chosen in the poselet selection step (described in section 3.3) collects 250 patches that have similar local joint configuration and uses them as positive examples for training. The patch size is set to one of 96 x 64, 64 x 64, 64 x 96 and 128 x 64 according to the aspect ratio of the area that covers the joints comprising a part. We use the distance metric $D(P1, P2) = D_{proc}(P1, P2) + \lambda D_{vis}(P1, P2)$ proposed by [2], where D_{proc} and D_{vis} are the Procrustes distance between joint configurations of both patches and a visibility distance which is set to the intersection over union

of joints present in both patches, respectively. We train a linear SVM classifier with positive examples and negative examples that are randomly selected from images which contain no person. We collect false positives with highest SVM scores as hard negatives (10 times as many as the number of positive examples) and retrain the linear SVM classifier. This process is iterated three times.

After training the poselets, we extract activations by a multi-scale sliding window scheme applied to the training images. Each activation is then labeled as a *true positive*, *false positive*, or *unknown*, using ground-truth annotations of people and their joints. For each training image, we determine matches between detections and ground-truth by comparing the detected bounding box to the ground-truth bounding box that encloses the ground-truth joints, as well as computing the Procrustes distance between the predicted joint locations (using the seed patch joint locations) and the ground-truth joint locations. Note that when computing the Procrustes distance, we exclude rotation because detecting by sliding window does not consider rotation. The latter labeling criterion, not used in [2], discards any false detection whose bounding box matches a ground-truth bounding box but whose associated joint locations are far from the ground-truth joint locations. Each activation which has an intersection over union with ground-truth more than 0.5 and whose Procrustes distance between joints is less than 0.3 is labeled as true positive. If the intersection over union with ground-truth is less than 0.1, the activation is labeled as a false positive for the purpose of the subsequent stages. Others remain unlabeled. Figure 3 shows some examples of activations labeled as true positives and unknown. Assuming that the joint distribution is Gaussian as in [2], the mean and variance of each joint are computed over true positive poselet activations, allowing each poselet to have an associated distribution over the position of joints.

3.3. Poselet Seed Selection

Our goal is to generate a set of poselets for each part that covers all appearance variations of that part over its configuration space. If we randomly choose poselet seeds and train on the nearest neighbors of those seeds as in [2, 4], we find that many of the training samples are not detected by the trained poselet (or by any other poselet), i.e., many of the training samples are not "covered" by the set of poselets. In addition to requiring that each training sample is covered by at least one poselet, we also require that the poselet covers at least one training sample that is not covered by any other poselet, otherwise the poselet would be redundant.

We introduce the poselet seed selection to generate an effective and efficient set of poselets by considering these two aspects. The poselet seed selection is an iterative process consisting of two steps: (i) *seed selection* and (ii) *set update*, and each step considers each aspect, respectively.

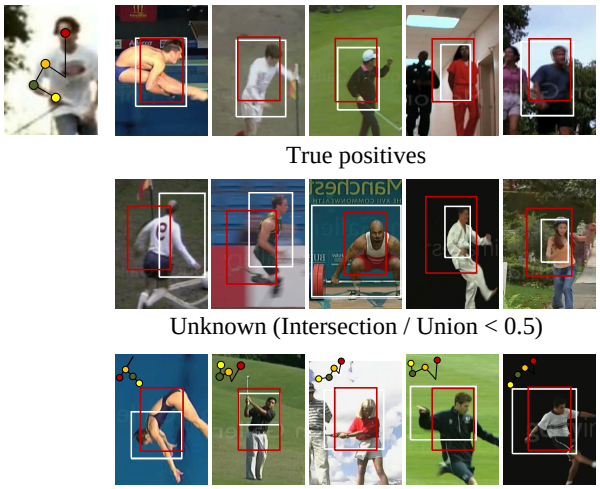

Figure 3. Examples of activations labeled as true positives and unknown. The top-left image shows a seed window for part 1 and a configuration of its joints. In the right column, 15 examples (5 for true positives, 10 for unknown activations) are shown in a right of the seed. White and red bounding boxes depict a groundtruth and detected window, respectively. In the third column, the configuration of its joints are depicated in a top-left corner of each image.

Denote that P is a set of poselets, and C is a list of training sample IDs that are covered by P. The set T of training patches is obtained from the physical joints annotated in the training set by enclosing the annotated joints with a bounding box (plus a suitable amount of padding). First, in the seed selection step, a patch not included in C is randomly selected and its poselet is trained. If a poselet is trained, example IDs containing any of its true positive activation are added to C. Second, the set update step identifies and removes poselets that are redundant (a poselet is redundant if all the patches it covers are already covered by other poselets). Given the coverage set C, a small size P is obtained by approximately solving a set cover problem, which is to identify the smallest subset which still covers all elements. We use a greedy algorithm to approximately solve the set cover problem. First, we sort all poselets in P in an ascending order according to the size of the subset covered by the poselets. Then, starting with the poselet with the smallest coverage, we remove any poselet from P if it is redundant.

3.4. Context Rescoring

After training the set of poselets to detect the root and the parts, we rescore activations by exploiting context among activations of the root and the parts. This step removes activation vectors that are not consistent with the detected human pose. We use labels of activations detected in training dataset for context rescoring. For

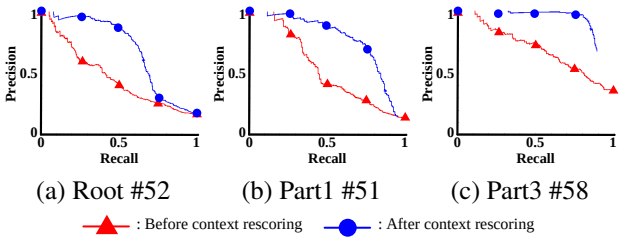

Figure 4. PR curves for performance of (a) root #52, (b) part 1 #51, and (c) part 3 #58 on YouTube action dataset. Red lines are obtained before context rescoring while blue lines are after context rescoring. Typical performance is shown for three randomly selected parts.

each root activation we obtain a set of consistent part activations, where consistency between root and part activation is measured by the symmetrized KL (Kullback-Liebler) divergence of their empirical joint distributions $d_{r,p} = \frac{1}{K}\sum_k D_{SKL}(N_r^k, N_p^k)$, where $D_{SKL}(N_r^k, N_p^k) = D_{KL}(N_r^k||N_p^k) + D_{KL}(N_p^k||N_r^k)$. Here, N_r^k and N_p^k are the empirical distributions of the k^{th} joint of root and part, respectively. We treat root and part as consistent if $d_{r,p}$ is below a threshold. For each root activation, we construct an activation vector consisting of the root poselet confidence score concatenated with a vector of the confidence scores of all part poselets. The score of the root activation is placed in the first bin and all consistent activations of parts are placed in their own bins according to the poselet type; multiple consistent activations of the same type are detected, but only the maximum score is entered in the appropriate bin. The remaining bins are filled with zero.

Then, we train a linear SVM classifier with activation vectors and their labels. At test time, root activations that are classified as false positives are discarded, and part activations with no mutually consistent root are also discarded as false positives. Figure 4 demonstrates that this context rescoring step effectively improves the precision-recall performance of both root and part poselet detectors by discarding many false positives; in the figure, root #52, part 1 #51, and part 3 #58 were arbitrarily chosen and have typical performance.

4. Video Representation

We extend the framework of [12] to include our proposed pose feature in addition to motion and shape features. For all features, initial histogram-based video representations are generated via bag-of-visual words (BoVW). After the initial representation is generated for each video sequence, compact yet discriminative visual vocabularies are obtained by feature grouping. A multi-class SVM classifier is trained using as input the concatenated visual word counts for each of the three features. Details about extracting motion, shape, and pose features are given in section 4.1 and the method for learning semantic visual vocabulary is described in section 4.2

4.1. Motion, Shape, and Pose Features

To complement our proposed pose feature, we select motion and shape features that achieve the best performances in [1, 12] on public datasets consisting of unconstrained videos.

Motion feature: We use the spatio-temporal interest point detector and descriptor proposed by Dollar et al. [5], which is described as being advantageous over other methods such as 3D Harris-Corner detector for action recognition in [12].

Shape feature: The shape feature uses the root position to compute a 3-level pyramid HOG around the root which shows the best performance among shape descriptors. [1] The region of interest side length is set to double the maximum value between the root's width and height.

Pose feature: We extract activations of root and parts by multi-scale sliding window and rescore root activations by context rescoring, using the activation vector constructed from all other mutually consistent poselet activations. Root activation vectors that are sufficiently confident after context rescoring (confidence > 0) are used as pose descriptors. The first bin in the activation vector corresponding to the root activation is excluded from the descriptor, since the root activation score is used only to confirm whether or not the root and consistent parts fit the particular qualitative pose model.

For each type of feature, we generate the histogram representation based on independent features via BoVW, which converts all features to "codewords" using k-means based on their descriptions.

4.2. Learning Semantic Visual Vocabulary

The initial vocabulary obtained by grouping similar features based on their appearance is far from semantically meaningful and its performance is sensitive to the size of the vocabulary, containing many redundant codewords that do not improve discrimination. We construct a compact yet discriminative visual vocabulary for each type of feature as proposed by [12]. A vocabulary is made compact by combining two bins of a BoVW if their class distributions are close to each other. Here, the distance between two distributions, p_1 and p_2 is measured by Jensen-Shannon (JS) divergence:

$$JS_\pi(p_1, p_2) = \sum_{i=1,2} \pi_i KL(p_i, \sum_{j=1,2} \pi_j p_j),$$
$$\pi_1 + \pi_2 = 1, \quad (1)$$

where $KL(\cdot)$ is the KL divergence.

Let $C = c_1, c_2, \cdots, c_L$ and $X = x_1, x_2, \cdots, x_M$ represent classes and codes, respectively. Let $\hat{X} =$

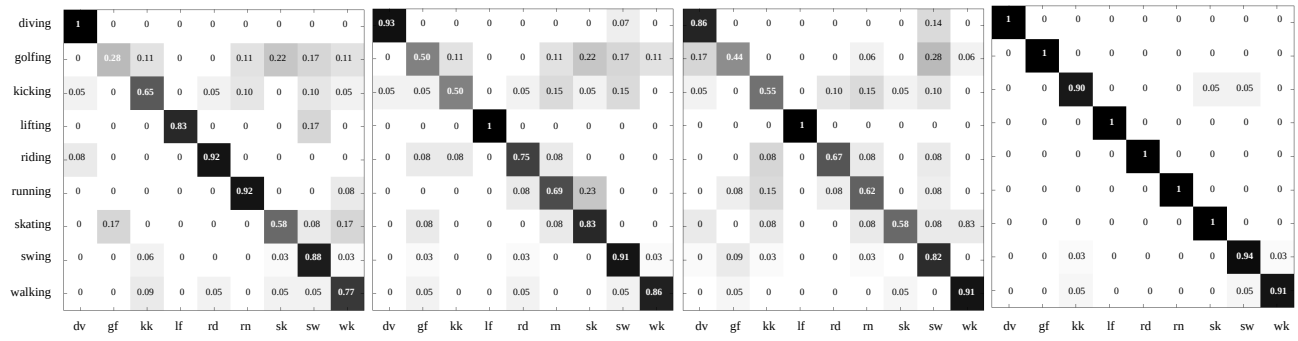

Figure 5. Confusion matrix for the YouTube sports [15] data set using combined feature with motion, pose, and shape feature.

$\hat{x}_1, \hat{x}_2, \cdots, \hat{x}_K$ be the updated clusters of X. A semantic visual vocabulary can be obtained by minimizing the loss of mutual information (MI), $Q(\hat{X}) = I(C;X) - I(C;\hat{X})$):

$$Q(\hat{X}) = \sum_{i=1}^{K} \pi(\hat{x}_i) JS(\{p(C|x_t) : x_t \in \hat{x}_i\}), \quad (2)$$

where $\pi(\hat{x}_i) = \sum_{x_t \in \hat{x}_t} \pi_t$, $\pi_t = p(x_t)$ is the prior. By equation 1, the mutual information is changed to

$$Q(\hat{X}) = \sum_{i=1}^{K} \pi(\hat{x}_i) \sum_{x_t \in \hat{x}_i} \pi_t KL(p(C|x_t), p(C|\hat{x}_i)). \quad (3)$$

The semantic representation $\hat{X}$ is generated by iterations of computing priors $\pi(\hat{x}_i), i = 1, 2, \cdots, K$ and updating clusters $i^*(x_t) = argmin_j KL(p(C|x_t), p(C|\hat{x}_i))$. A termination condition of the iteration is $Q(\hat{X}) < \epsilon$.

5. Experiments

We evaluate our framework on two benchmarks: YouTube sports dataset [15] and YouTube action dataset [12]. For both datasets, we follow the original authors' setting for evaluation. The multi-class linear SVM is used as the classifier for action recognition with vectors combining semantic representations of motion, pose, and shape feature. Each feature is normalized by L2 norm. Finally, we evaluate the boost in performance provided by our proposed poselet seed selection versus the original scheme proposed in [2]. All clustering parameters, including the size of the initial and semantic vocabulary, are obtained automatically by cross validation.

5.1. Experiments on YouTube Sports Dataset

The YouTube sports dataset [15] consists of a set of actions collected from various sports which are typically seen in broadcast media. For each feature, we set the initial vocabulary size to 500 and the semantic vocabulary size to 100. During training, we store for each poselet the video sequence from which its training images were selected. For clustering, we set the portion of coverage to 0.8, resulting in 123, 120, 120, and 123 poselets for the root and the three parts, respectively.

Method	Accuracy (%)
Wang et al. [18]	85.6
Le et al. [10]	86.5
O'Hara and Draper [14]	91.3
Todorovic [17]	92.1
Sadanand and Corso [16]	95.0
Shape	71.3
Motion	75.3
Pose	76.7
Pose + Shape	84.7
Motion + Shape	86.7
Motion + Pose	90.7
Motion + Pose + Shape	**96.0**

Table 2. Recognition rates on the YouTube sports data set.

Figure 5 shows the confusion matrix for classification using motion, pose, shape, and hybrid (combination of all three) features. The motion feature is useful for classifying actions in which human locations change significantly, e.g., diving, horseback riding, and running. On the other hand, the pose feature outperforms others for actions consisting of distinctive poses, e.g., arm pose after golf swing or lifting and pose of legs when skating. For walking, the shape feature yields the best classification performance since walking does not involve particularly distinctive motions or poses. In table 2, the recognition rates using pose feature are the highest among the three types of features. Using a hybrid of motion, pose, and shape features yields an improvement in performance over Sadanand and Corso [16], the state-of-the-art.

5.2. Experiments on YouTube Action Dataset

We also evaluate our framework on the challenging YouTube action dataset [12] consisting of 11 action classes.

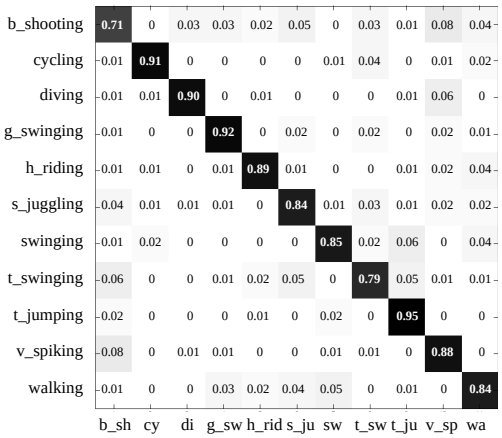

Figure 6. Confusion matrix for the YouTube action [12] data set using combined feature with motion, pose, and shape feature.

Method	Accuracy (%)
Liu et al. [12]	71.2
Zhang et al. [23]	72.9
Le et al. [10]	75.8
Shape	52.3
Motion	62.2
Motion + Shape	72.9
Pose	74.6
Pose + Shape	76.0
Motion + Pose	83.5
Motion + Pose + Shape	**86.2**

Table 3. Recognition rates on the YouTube action data set. We outperform the state-of-the-art by over 10%.

For clustering, we select 100 poselets for the root and each part. Here, we set the size of the initial vocabulary and semantic vocabulary to 1000 and 100, respectively.

Figure 6 shows the confusion matrix for the YouTube action dataset. Based on the confusion matrix, our framework has the worst performance on basketball shooting and walking. Because the pose observed during shooting in basketball is similar to swinging an arm in tennis or spiking in volleyball, most of the miss-classified video sequences are classified into those classes. The reason for the low classification performance for walking is likely the same as for the previous dataset. In table 3, our framework outperformed other algorithms by approximately 10.4%. Interestingly, using pose feature alone provides recognition rates which matches all the state-of-the-art. Figure 7 shows some examples of pose features for a qualitative evaluation.

5.3. Boost by Poselet Seed Selection

In this section, we compare our proposed poselet seed selection process against the random selection process of [2] in performance. The proposed selection process results in a set of poselets that cover 80% of the training examples (a training sample is *covered* if the poselet detector yields an activation that overlaps sufficiently with the training sample), which results in a final recognition rate of 76.7 on the youtube sports dataset [15]. The sizes of the poselets set for root, part 1, and part 3 are 123, 120, and 123, respectively. Part 1 and 2 are mirrored versions of each other, thus yielding a total of 486 poselets. Table 4 shows the performance over various numbers of poselets chosen by random selection versus our approach. As the number of poselet grows, the coverage of the training dataset and recognition rate improves but does not match the recognition rate obtained by our proposed poselet seed selection until training reaches 300 poselets for the root and each part (for a total of 1200 for the root and the three parts, as in [3]).

		random selection			proposed
number of poselets		400	800	1200	486
covered set (%)	root	56.1	60.6	62.3	**80.1**
	part 1	48.7	54.4	56.5	**80.0**
	part 3	53.3	59.3	61.7	**80.1**
recognition rate (%)		63.3	67.3	71.3	**76.7**

Table 4. Top rows: the percentage of the training dataset *covered* (see text) as the number of total poselets is varied. Bottom row: the resulting action recognition rates. The right column shows the coverage and recognition rates of our proposed selection approach.

6. Conclusion

We proposed a robust pose feature based on poselets that is suitable for use in action recognition tasks involving relatively unconstrained videos. We have shown that various modifications of the poselet training process improve the representation power of the set of poselets, generating a set of features that can be seamlessly combined with existing shape and motion features. Experiments show that our proposed pose feature provides significant information alone; when in addition to motion and shape, we obtain state-of-the-art results.

References

[1] A. Bosch, A. Zisserman, and X. Munoz. Representing shape with a spatial pyramid kernel. In *CIVR*, 2007.

[2] L. Bourdev, S. Maji, T. Brox, and J. Malik. Detecting people using mutually consistent poselet activations. In *ECCV*, 2010.

[3] L. Bourdev, S. Maji, and J. Malik. Describing people: Poselet-based approach to attribute classification. In *ICCV*, 2011.

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

[5] P. Dollar, V. Raboud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatiotemporal features. In *VS-PETS*, 2005.

Figure 7. Example root and part activations for each class in the YouTube action dataset. The left-most image in each example is a region of the test image cropped by the root activation bounding box (plus padding), with consistent parts highlighted. The average image of some detected poselet is shown to the right (note: there are other activations that are not shown due to space constraints).

[6] R. Farrell, O. Oza, N. Zhang, V. Morariu, T. Darrell, and L. Davis. Subordinate categorization using volumetric primitives and pose-normalized appearance. In *ICCV*, 2011.

[7] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *PAMI*, 32(9), 2010.

[8] Z. Jiang, Z. Lin, and L. S. Davis. Recognizing human actions by learning and matching shape-motion prototype trees. *PAMI*, 34(3), 2012.

[9] I. Laptev and T. Lindeberg. Space-time interest points. In *ICCV*, 2003.

[10] Q. Le, W. Z. S. Yeung, and A. Ng. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In *CVPR*, 2011.

[11] H. Lee, V. I. Morariu, and L. S. Davis. Qualitative pose estimation by discriminative deformable part models. In *ACCV*, 2012.

[12] J. Liu, J. Luo, and M. Shah. Recognizing realistic actions from videos "in the wild". In *CVPR*, 2009.

[13] S. Maji, L. Bourdev, and J. Malik. Action recognition from a distributed representation of pose and appearance. In *CVPR*, 2011.

[14] S. O'Hara and B. A. Draper. Scalable action recognition with a subspace forest. In *CVPR*, 2012.

[15] M. D. Rodriguez, J. Ahmed, and M. Shah. Action mach: A spatio-temporal maximum average correlation height filter for action recognition. In *CVPR*, 2008.

[16] S. Sadanand and J. J. Corso. Action bank: A high-level representation of activity in video. In *CVPR*, 2012.

[17] S. Todorovic. Human activities as stochastic kronecker graphs. In *ECCV*, 2012.

[18] H. Wang, M. M. Ullah, A. Kläser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In *BMVC*, 2009.

[19] G. Willems, T. Tuytelaars, and L. V. Gool. An efficient dense and scale-invariant spatio-temporal interest point detector. In *ECCV*, 2008.

[20] Y. Xie, H. Chang, Z. Li, L. Liang, X. Chen, and D. Zhao. A unified framework for locating and recognizing human actions. In *CVPR*, 2011.

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

[22] B. Yao, X. Jiang, A. Khosla, A. L. Lin, L. Guibas, and L. Fei-Fei. Human action recognition by learning bases of action attributes and parts. In *ICCV*, 2011.

[23] Y. Zhang, X. Liu, M.-C. Chang, W. Ge, and T. Chen. Spatio-temporal phrases for activity recognition. In *ECCV*, 2012.