Human Action Recognition by Learning Bases of Action Attributes and Parts

Bangpeng Yao[1], Xiaoye Jiang[2], Aditya Khosla[1], Andy Lai Lin[3], Leonidas Guibas[1], and Li Fei-Fei[1]

[1]Computer Science Department, Stanford University, Stanford, CA
[2]Institute for Computational & Mathematical Engineering, Stanford University, Stanford, CA
[3]Electrical Engineering Department, Stanford University, Stanford, CA

{bangpeng,aditya86,guibas,feifeili}@cs.stanford.edu {xiaoye,ydna}@stanford.edu

Abstract

In this work, we propose to use attributes *and* parts *for recognizing human actions in still images. We define action attributes as the verbs that describe the properties of human actions, while the parts of actions are objects and poselets that are closely related to the actions. We jointly model the attributes and parts by learning a set of sparse bases that are shown to carry much semantic meaning. Then, the attributes and parts of an action image can be reconstructed from sparse coefficients with respect to the learned bases. This* dual sparsity *provides theoretical guarantee of our bases learning and feature reconstruction approach. On the PASCAL action dataset and a new "Stanford 40 Actions" dataset, we show that our method extracts meaningful high-order interactions between attributes and parts in human actions while achieving state-of-the-art classification performance.*

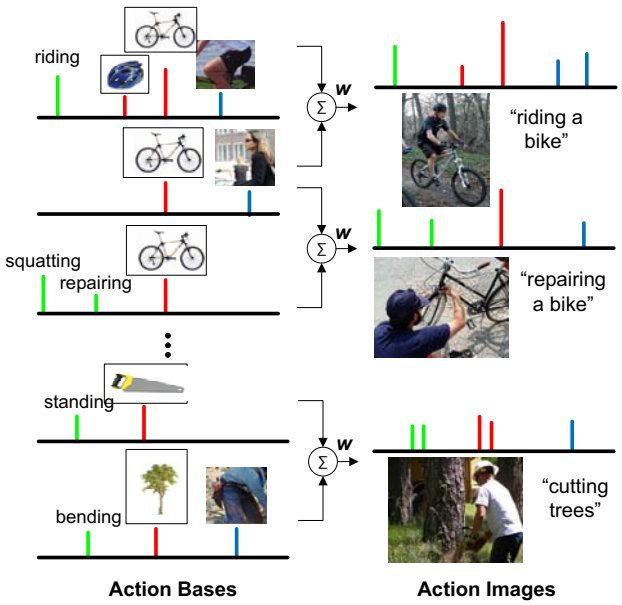

Figure 1. We use attributes (verb related properties) and parts (objects and poselets [2]) to model action images. Given a large number of image attributes and parts, we learn a number of sparse action bases, where each basis encodes the interactions between some highly related attributes, objects, and poselets. The attributes and parts of an image can be reconstructed from a sparse weighted summation of those bases. The colored bars indicate different attributes and parts, where the color code is: green - attribute, red - object, blue - poselet. The height of a bar reflects the importance of this attribute or part in the corresponding basis.

1. Introduction

Recognizing human actions in still images has many potential applications in image indexing and retrieval. One straightforward solution for this problem is to use the whole image to represent an action and treat action recognition as a general image classification problem [13, 28, 4, 30]. Such methods have achieved promising performance on the recent PASCAL challenge using spatial pyramid [16, 4] or random forest [30] based methods. These methods do not, however, explore the semantically meaningful components of an action, such as human poses and the objects that are closely related to the action.

There is some recent work which uses objects [12, 29, 6, 24] interacting with the person or human poses [27, 22] to build action classifiers. However, these methods are prone to problems caused by false object detections or inaccurate pose estimations. To alleviate these issues, some methods [29] rely on labor-intensive annotations of objects and human body parts during training time, posing a serious concern towards large scale action recognition.

Inspired by the recent work on using objects and body parts for action recognition as well as global and local attributes [9, 15, 1, 23] for object recognition, in this paper, we propose an *attributes* and *parts* based representation of human actions in a weakly supervised setting. The action attributes are holistic image descriptions of human actions,

usually associated with verbs in the human language such as "riding" and "sitting" (as opposed to "repairing" or "lifting") for the action "riding bike". The action parts include objects that are related to the corresponding action (e.g. "bike", "helmet", and "road" in "riding bike") as well as different configurations of local body parts (we use poselet described in [2]). Given an image of a human action, many attributes and parts[1] contribute to the recognition of the corresponding action.

Given an image collection of many different actions, there is a large number of possible attributes, objects and poselets. Furthermore, there is a large number of possible interactions among these attributes and parts in terms of co-occurrence statistics. For example, the "riding" attribute is likely to co-occur with objects such as "horse" and "bike", but not "laptop", while the "right arm extended upward" poselet is more likely to co-occur with objects such as "volleyball" and the attribute "hitting". We formulate these interactions of action attributes and parts as *action bases* for expressing human actions. A particular action in an image can therefore be represented as a weighted summation of a subset of these bases, as shown in Fig.1.

This representation can be naturally formulated as a reconstruction problem. Our challenge is to: 1) represent each image by using a sparse set of action bases that are meaningful to the content of the image, 2) effectively learn these bases given far-from-perfect detections of action attributes and parts without meticulous human labeling as proposed in previous work [29]. To resolve these challenges, we propose a *dual sparsity* reconstruction framework to simultaneously obtain sparsity in terms of both the action bases as well as the reconstruction coefficients for each image. We show that our method has theoretical foundations in sparse coding and compressed sensing [32, 14]. On the PASCAL action dataset [7] and a new "Stanford 40 Actions" dataset, our attributes and parts representation significantly outperforms state-of-the-art methods. Furthermore, we visualize the bases obtained by our framework and show semantically meaningful interpretations of the images.

The remaining part of this paper is organized as follows. Related work are described in Sec.2. The attributes and parts based representation of actions and the method to learn action bases are elaborated in Sec.3 and Sec.4 respectively. Experiment results are shown and discussed in Sec.5.

2. Related Work

Most of the action recognition approaches [28, 4, 7] for still images treat the problem as a pure image classification problem. There are also algorithms which model the objects or human poses for action classification, such as the mutual context model [29] and poselets [2, 22]. However, the mutual context model requires supervision of the bounding boxes of objects and human body parts, which are expensive to obtain especially when there is a large number of images. Also, we want to put the objects and human poses in a more discriminative framework so that the action recognition performance can be further improved. While poselets have achieved promising performance on action recognition [22], it is unclear how to jointly explore the semantic meanings of poselets and the other concepts such as objects for action recognition.

In this paper, we propose to use attributes and parts for action classification. Inspired by the recent work of learning attributes for object recognition [9, 15, 1, 23] and action recognition in videos [19], the attributes we use are linguistically related description of the actions. We use a global image based representation to train a classifier for each attribute. Compared to the attributes for objects which are usually adjectives or shape related, the attributes we use to describe actions are mostly related to verbs. The parts based models have been successfully used in object detection [10] and recognition [11]. However unlike these approaches that use low-level descriptors, the action parts we use are objects and poselets with pre-trained detectors as in [18, 22]. The discriminative information in those detectors can help us alleviate the problem of background clutter in action images and give us more semantic information of the images [18].

In the attributes and parts based representation, we learn a set of sparse action bases and estimate a set of coefficients on these bases for each image. This dual sparsity makes our problem different from traditional dictionary learning and sparse coding problems [25, 17, 21], given that our action bases are sparse (in the large set of attributes and parts, only a small number of them are highly related in each basis) and far from being mutually orthogonal (consider the two bases "riding - sitting - bike" and "riding - sitting - horse"). In this work, we solve this dual sparsity problem using the elastic-net constrained set [32], and show that our approach has theoretical foundations in the compressed network theorem [14].

3. Action Recognition with Attributes & Parts

3.1. Attributes and Parts in Human Actions

Our method jointly models different attributes and parts of human actions, which are defined as follows.

Attributes: The attributes are linguistically related descriptions of human actions. Most of the attributes we use are related to verbs in human language. For example, the attributes for describing "riding a bike" can be "riding" and "sitting (on a bike seat)". It is possible for one attribute to correspond to more than one action. For instance, "riding"

[1] Our definition of action attributes and parts are different from the attributes and parts in common object recognition literature. Please refer to Sec.2 for details. In this work we use "action attribute" and "attribute", "action part" and "part" interchangeably, if not explicitly specified.

can describe both "riding a bike" and "riding a horse", while this attribute can differentiate the intentions and human gestures in the two actions with the other ones such as "drinking water". Inspired by the previous work on attributes for object recognition [9, 15, 1], we train a discriminative classifier for each attribute.

Parts: The parts we use are composed of objects and human poses. We assume that an action image consists of the objects that are closely related to the action and the descriptive local human poses. The objects are either manipulated by the person (e.g. "bike" in "riding a bike") or related to the scene context of the action (e.g. "road" in "riding a bike", "reading lamp" in "reading a book"). The human poses are represented by poselets [2], where the human body parts in different images described by the same poselet are tightly clustered in both appearance space and configuration space. In our approach, each part is modeled by a pre-trained object detector or poselet detector.

To obtain our features, we run all the attribute classifiers and part detectors on a given image. A vector of the normalized confidence scores obtained from these classifiers and detectors is used to represent this image.

3.2. Action Bases of Attributes and Parts

Our method learns high-order interactions of image attributes and parts. Each interaction corresponds to the co-occurrence of a set of attributes and parts with some specific confidence values (Fig.1). These interactions carry richer information about human actions and are thus expected to improve recognition performance. Furthermore, the components in each high-order interaction can serve as context for each other, and therefore the noise in the attribute classifiers and part detectors can be reduced. In our approach, the high-order interactions are regarded as the bases of the representations of human actions, and each image is represented as a sparse distribution with respect to all the bases. Examples of the learned action bases are shown in Fig.4. We can see that the bases are sparse in the whole space of attributes and parts, and many of the attributes and parts are closely correlated in human actions, such as "riding - sitting - bike" and "using - keyboard - monitor - sitting" as well as the corresponding poselets.

Now we formalize the action bases in a mathematical framework. Assume we have P attributes and parts, and let $\mathbf{a} \in \mathbb{R}^P$ be the vector of confidence scores obtained from the attribute classifiers and part detectors. Denoting the set of action bases as $\mathbf{\Phi} = [\boldsymbol{\phi}_1, \cdots, \boldsymbol{\phi}_M]$ where each $\boldsymbol{\phi}_m \in \mathbb{R}^P$ is a basis, the vector $\mathbf{a}$ can be represented as

$$\mathbf{a} = \sum_{m=1}^{M} w_m \boldsymbol{\phi}_m + \boldsymbol{\varepsilon} \tag{1}$$

where $\mathbf{w} = \{w_1, \cdots, w_M\}$ are the reconstruction coefficients of the bases, and $\boldsymbol{\varepsilon} \in \mathbb{R}^P$ is a noise vector. Note that in our problem, the vector $\mathbf{w}$ and $\{\boldsymbol{\phi}_m\}_{m=1}^M$ are all sparse. This is because on one hand, only a small number of attributes and parts are highly related in each basis of human actions; on the other hand, a small proportion of the action bases are enough to reconstruct the set of attributes and parts in each image.

3.3. Action Classification Using the Action Bases

From Eqn.1, we can see that the attributes and parts representation $\mathbf{a}$ of an action image can be reconstructed from the sparse factorization coefficients $\mathbf{w}$. $\mathbf{w}$ reflects the distribution of $\mathbf{a}$ on all the action bases $\mathbf{\Phi}$, each of which encodes a specific interaction between action attributes and parts. The images that correspond to the same action should have high coefficients on the similar set of action bases. In this paper, we use the coefficients vector $\mathbf{w}$ to represent an image, and train an SVM classifier for action classification.

The above classification approach resolves the two challenges of using attributes and parts (objects and poselets) for action recognition that we proposed in Sec.1. Since we only use the learned action bases to reconstruct the feature vector, our method *can correct some false detections of objects and poselets* by removing the noise component ε in Eqn.1. Also, those action bases correspond to some high-order interactions in the features, and therefore they *jointly model the complex interactions between different attributes, objects, and poselets*.

4. Learning the Dual-Sparse Action Bases and Reconstruction Coefficients

Given a collection of training images represented as $\mathcal{A} = \{\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_N\}$ as described in Sec.3.2, where each $\mathbf{a}_i$ is the vector of confidence scores of attribute classifications and part detections computed from image i. Intuitively, there exists a latent dictionary of bases where each basis characterizes frequent co-occurrence of attributes, objects, and poselets involved in an action, e.g. "cycling" and "bike", such that each observed data $\mathbf{a}_i$ can be sparsely reconstructed with respect to the dictionary. Our goal is to identify a set of sparse bases $\mathbf{\Phi} = [\boldsymbol{\phi}_1, \cdots, \boldsymbol{\phi}_M]$ such that each $\mathbf{a}_i$ has a sparse representation with respect to the dictionary, as shown in Eqn.1.

During the bases learning stage, we need to learn the bases $\mathbf{\Phi}$ and find the reconstruction coefficients $\mathbf{w}_i$ for each $\mathbf{a}_i$. Given a new image represented by $\mathbf{a}$, we want to find a sparse $\mathbf{w}$ such that $\mathbf{a}$ can be reconstructed from the learned $\mathbf{\Phi}$. Therefore our bases learning and action reconstruction can be achieved by the following two optimization problems respectively,

$$\min_{\mathbf{\Phi} \in \mathcal{C}, \mathbf{W} \in \mathbb{R}^{M \times N}} \sum_{i=1}^{N} \left(\frac{1}{2} \|\mathbf{a}_i - \mathbf{\Phi}\mathbf{w}_i\|_2^2 + \lambda \|\mathbf{w}_i\|_1 \right), \tag{2}$$

$$\min_{\mathbf{w}\in\mathbb{R}^M} \frac{1}{2}\|\mathbf{a}-\mathbf{\Phi}\mathbf{w}\|_2^2+\lambda\|\mathbf{w}\|_1, \quad (3)$$

where $\mathbf{W}=[\mathbf{w}_1,\cdots,\mathbf{w}_N]\in\mathbb{R}^{M\times N}$, λ is a regularization parameter, and $\mathcal{C}$ is the convex set that $\mathbf{\Phi}$ belongs to. The l_1-norm in Eqn.2 makes the reconstruction coefficients $\mathbf{w}_i$ tend to be sparse. In our setting, the bases $\mathbf{\Phi}$ should also be sparse, even though the given $\mathcal{A}$ might be quite noisy due to the error-prone object detectors and poselet detectors. To address this issue, we construct the convex set $\mathcal{C}$ as:

$$\mathcal{C}=\{\mathbf{\Phi}\in\mathbb{R}^{P\times M}, \text{s.t. } \forall j,\ \|\mathbf{\Phi}_j\|_1+\frac{\gamma}{2}\|\mathbf{\Phi}_j\|_2^2\le 1\}. \quad (4)$$

where γ is another regularization parameter.

Including both l_1-norm and l_2-norm to define the convex set $\mathcal{C}$, the sparsity requirement of the bases are encoded. This is called the *elastic-net constraint set* [32]. Furthermore, the sparsity on $\mathbf{\Phi}$ implies that different action bases have small overlaps, therefore the coefficients learned from Eqn.2 are guaranteed to generalize to the testing case in Eqn.3 according to the compressed network theorem [14]. Please refer to the supplementary document[2] for details.

In our two optimization problems, Eqn.3 is convex while Eqn.2 is non-convex. However Eqn.2 is convex with respect to each of the two variables $\mathbf{\Phi}$ and $\mathbf{W}$ when the other one is fixed. We use an online learning algorithm [21] which scales up to large datasets to solve this problem.

5. Experiments and Results

5.1. Dataset and Experiment Setup

We test the performance of our proposed method on the PASCAL action dataset [7] and a new larger scale dataset collected by us. The new dataset, called *Stanford 40 Actions*, contains 40 diverse daily human actions, such as "brushing teeth", "cleaning the floor", "reading book", "throwing a frisbee", etc. All the images are obtained from Google, Bing, and Flickr. We collect 180∼300 images for each class. The images within each class have large variations in human pose, appearance, and background clutter. The comparison between our dataset and the existing still image action datasets are summarized in Table 1. As there might be multiple people in a single image, we provide bounding boxes for the humans who are doing one of the 40 actions in each image, similar to [7]. Examples of the images in our dataset[3] are shown in Fig.2.

On the PASCAL dataset, we use the training and validation set specified in [7] for training, and use the same testing set. On the Stanford 40 Action dataset, we randomly select 100 images in each class for training, and the remaining images for testing. For each dataset, we annotate the attributes that can be used to describe the action in each image, and then train a binary classifier for each attribute. We take a global representation of the attributes as in [9], and use the Locality-constrained Linear Coding (LLC) method [26] on dense SIFT [20] features to train the classifier for each attribute. As in [4], the classifiers are trained by concatenating the features from both the foreground bounding box of the action and the whole image. We extend and normalize the bounding boxes in the same way as in [4]. For objects, we use the ImageNet [5] dataset with provided bounding boxes to train the object detectors by using the Deformable Parts Model [10], instead of annotating the positions of objects in

Figure 2. Example images of the Stanford 40 Actions Dataset.

Dataset	No. of actions	No. of images	Clutter?	Poses vary?	Visibility varies?
Ikizler [13]	5	1727	**Yes**	**Yes**	**Yes**
Gupta [12]	6	300	Small	Small	No
PPMI [28]	**24**	**4800**	**Yes**	**Yes**	No
PASCAL [7]	9	1221	**Yes**	**Yes**	**Yes**
Stanford 40	**40**	**9532**	**Yes**	**Yes**	**Yes**

Table 1. Comparison of our Stanford 40 Action dataset and other existing human action datasets on still images. "Visibility" variation refers to the variation of visible human body parts, e.g. in some images the full human body is visible, while in some other images only the head and shoulder are visible. Bold font indicate relatively larger scale datasets or larger image variations.

[2]The supplementary document can be found on the author's website.

[3]Please refer to http://vision.stanford.edu/Datasets/40actions.html for more details of the Stanford 40 Actions dataset.

Method	Phoning	Playing instrument	Reading	Riding bike	Riding horse	Running	Taking photo	Using computer	Walking	Overall
SURREY_MK	**52.6**	53.5	35.9	81.0	89.3	86.5	32.8	59.2	68.6	62.2
UCLEAR_DOSP	47.0	57.8	26.9	78.8	89.7	87.3	32.5	60.0	70.1	61.1
WILLOW_LSVM	49.2	37.7	22.2	73.2	77.1	81.7	24.3	53.7	56.9	52.9
POSELETS	45.9	45.8	23.7	79.9	87.6	83.1	26.2	44.9	66.6	56.0
Ours Conf_Score	49.5	56.6	31.4	**82.3**	89.3	87.0	36.1	**67.7**	73.0	63.7
Ours Sparse_Bases	42.8	**60.8**	**41.5**	80.2	**90.6**	**87.8**	**41.4**	66.1	**74.4**	**65.1**

Table 2. Comparison of our method and the other action classification approaches evaluated using the percentage of average precision. "Overall" indicates the mean Average Precision (mAP) on all the nine classes. The bold fonts indicate the best performance. SURREY_MK UCLEAR_DOSP, WILLOW_SVMSIFT, and POSELETS are the approaches presented in the PASCAL challenge [7].

the action data. For poselets, we use the pre-trained poselet detectors in [2]. For each object or poselet detector, we use the highest detection score in the response map of each image to measure the confidence of the object or poselet in the given image. We linearly normalize the confidence scores of all the attribute classifiers and part detectors so that all the feature values are between 0 and 1.

We use 14 attributes and 27 objects for the PASCAL data, 45 attributes and 81 objects for the Stanford 40 Action data. We only use the attributes and objects that we believe are closely related to the actions in each dataset. Also some useful objects are not included, e.g. cigarette which is helpful for recognizing the action of "smoking cigarette" but there is no cigarette bounding box in ImageNet images. Please refer to the supplementary document for the list of attributes and objects that we use. We use 150 poselets as provided in [2] on both datasets. The number of action bases are set to 400 and 600 respectively. The λ and γ values in Eqn.2, 3, and 4 are set to 0.1 and 0.15.

In the following experiment, we consider two approaches of using attributes and parts for action recognition. One is to simply concatenate the normalized confidence scores of attributes classification and parts detection as feature representation (denoted as "Conf_Score"), the other is to use the reconstruction coefficients on the learned sparse bases as feature representation (denoted as "Sparse_Bases"). We use linear SVM classifiers for both feature representations. As in [7], we use mean Average Precision (mAP) to evaluate the performance on both datasets.

5.2. Results on the PASCAL Action Dataset

On the PASCAL dataset, we compare our methods with four approaches from the PASCAL challenge [7]: the SURREY_MK and UCLEAR_DOSP which mainly rely on general image classification methods and achieve the best performance in the challenge, WILLOW_LSVM which is a parts based model, and POSELETS which also uses the poselet features for classification.

The average precision of different approaches is shown in Table 2. We can see that by simply concatenating the confidence scores of attributes classification and parts detection, our method outperforms the best result in the PASCAL challenge in terms of the mean Average Precision (mAP). The performance can be further improved by learning high-order interactions of attributes and parts, from which the feature noise can be reduced. A visualization of the learned bases of our method is shown in Fig.4. We observe that almost all the bases are very sparse, and many of them carry useful information for describing specific human actions. However due to the large degree of noise in both object detectors and poselet detectors, some bases contain noise, e.g. "guitar" in the basis of "calling - cell phone - guitar". In Fig.4 we also show some action images with the annotations of attributes and objects that have high confidence score in the feature representation reconstructed from the bases.

Our approach considers three concepts: attributes, parts as objects, and parts as poselets. To analyze the contribution of each concept, we remove the confidence scores of attribute classifiers, part detectors, and poselet detectors from our feature set, one at a time. The classification results are shown in Fig.3. We observe that using the reconstruction coefficients consistently outperform the methods that simply concatenating the confidence scores of classifiers and

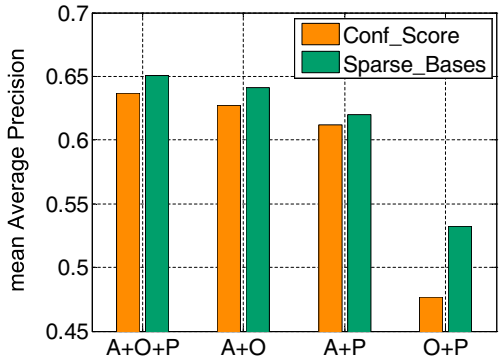

Figure 3. Comparison of the methods by removing the confidence scores obtained from attributes (A), objects (O), and poselets (P) from the feature vector, one at a time. The performance are evaluated using mean Average Precision on the PASCAL dataset.

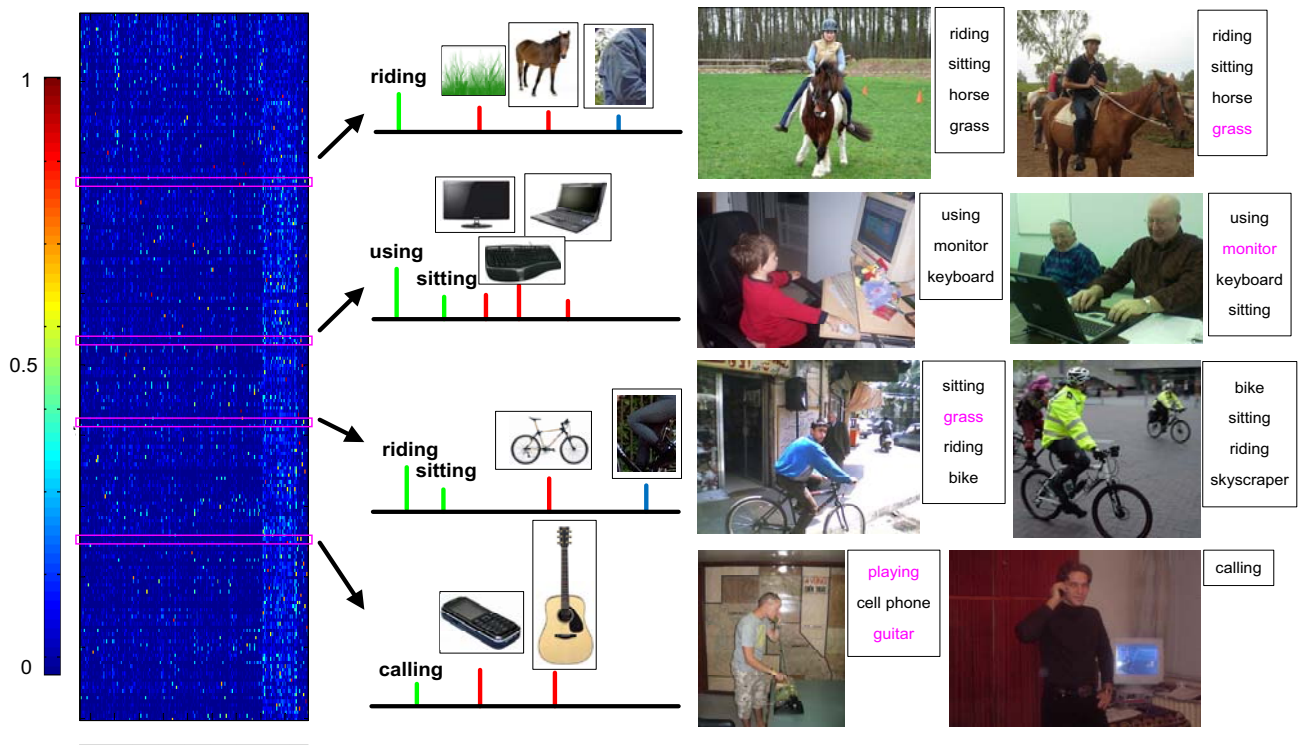

Figure 4. Visualization of the 400 learned bases from the PASCAL action dataset. Each row in the left-most matrix corresponds to one basis. Red color indicates large magnitude in the action bases while blue color indicates low magnitude. We observe that the bases are indeed very sparse. We also show some semantically meaningful action bases learned by our results, e.g. "riding - grass - horse". By using the learned action bases to reconstruct the attributes and parts representation, we show the attributes and objects that have high confidence scores on some images. Magenta color indicates wrong tags.

detectors. We can also see that attributes make the biggest contribution to the performance, because removing the attribute features makes the performance much worse. This is due to the large amount of noise produced from objects and poselets detectors which are pre-trained from the other datasets. However, objects and poselets do contain complementary information with the attributes, and the effect of the noise can be alleviated by the bases learned from our approach. We observe that in the case of only considering objects and poselets, learning the sparse bases significantly improves the performance. By combining attributes, objects and poselets and learning the action bases, our method achieves state-of-the-art classification performance.

Our learning method (Eqn.2) has the dual sparsity on both action bases $\mathbf{\Phi}$ and reconstruction coefficients $\mathbf{W}$. Here we compare our method with a simple l_1-norm method - l_1 logistic regression based on the concatenation of the confidence scores of attributes and parts. The mAP result of l_1 logistic regression is 47.9%, which is lower than our results. This shows that a simple l_1-norm logistic regression cannot effectively learn the information from the noisy attributes classification and parts detection features. Furthermore, in order to demonstrate the effectiveness of the two sparsity constraints, we remove the constraints one at a time. To remove the sparsity constraint on the reconstruction weight $\mathbf{W}$, we simply change $\|\mathbf{w}_i\|_1$ in Eqn.2 and Eqn.3 to $\|\mathbf{w}_i\|_2$. To remove the sparsity constraint on the bases $\mathbf{\Phi}$, we change the convex set $\mathcal{C}$ in Eqn.4 to be:

$$\mathcal{C} = \{\mathbf{\Phi} \in \mathbb{R}^{P\times M}, \text{s.t. } \forall j,\ \|\mathbf{\Phi}_j\|_2^2 \leq 1\}. \tag{5}$$

In the first case, where we do not have sparsity constraint on $\mathbf{W}$, the mAP result drops to 64.0%, which is comparable to directly concatenating all attributes classification and parts detection confidence scores. This shows that the sparsity on $\mathbf{W}$ helps to remove noise from the original data. In the second case where we do not have sparsity constraint on $\mathbf{\Phi}$, the performance becomes 64.7% which is very close to that of having sparsity constraint on $\mathbf{\Phi}$. The reason might be that although there is much noise in the parts detections and attribute classifications, the original vector of confidence scores already has some level of sparsity. However, by explicitly imposing the sparsity on $\mathbf{\Phi}$, we can guarantee the sparsity of the bases, so that our method can explicitly extract more semantic information and its performance is also theoretically guaranteed. Please refer to the supplementary document of this paper for more details.

5.3. Results on the Stanford 40 Actions Dataset

We next show the performance of our proposed method on the new Stanford 40 Actions dataset. We setup two baselines on this dataset: LLC [26] method with densely sampled SIFT [4] features, and object bank [18]. Comparing these two algorithms with our approach, the mAP is shown in Table 3. The results show that compared to the baselines which uses image classifiers or object detectors only, combining attributes and parts (objects and poselets) significantly improved the recognition performance by more than 10%. The reason might be that, on this relatively large dataset, more attributes are used to describe the actions and more objects are related to the actions, which contains a lot of complementary information.

As done in Sec.5.2, we also remove the features that are related to attributes, objects, and poselets from our feature set, one at a time. The results are shown in Fig.5. On this dataset, the contribution of objects is larger than that on the PASCAL dataset. This is because more objects are related to the actions on this larger scale dataset, and therefore we can extract more useful information for recognition from the object detectors.

The average precision obtained from LLC and our method by using reconstruction coefficients as feature representation for each of the 40 classes is shown in Fig.6. Using a sparse representation on the action bases of attributes and parts, our method outperforms LLC on all the 40 classes. Furthermore, the classification performance on different actions varies a lot, ranging from 89.2% on "riding a horse" to only 6.2% on "texting message". It is interesting to observe that the result shown in Fig.6 is somewhat similar to that on the PASCAL dataset in Table 2. The classes "riding a horse" and "riding a bike" have high classification performance on both datasets while the classes "calling", "reading a book" and "taking photos" have low classification performance, showing that the two datasets capture similar image statistics of human actions. The classes "riding a horse" and "riding a bike" can be easily recognized in part because the human poses do not vary much within each action, and the objects (horse and bike) are easy to detect. However, the performance on "feeding a horse" and "repairing a bike" is not as good as that on "riding a horse" and "riding a bike". One reason is that the body parts of horses in most of the images of "feeding a horse" are highly

Method	Object Bank [18]	LLC [26]	Ours Conf_Score	Ours Sparse_Bases
mAP	32.5%	35.2%	44.6%	**45.7%**

Table 3. Comparison of our attributes and parts based action recognition methods with the two baselines: object bank [18] and LLC [26]. The performance is evaluated with AP. The bold font indicates the best performance.

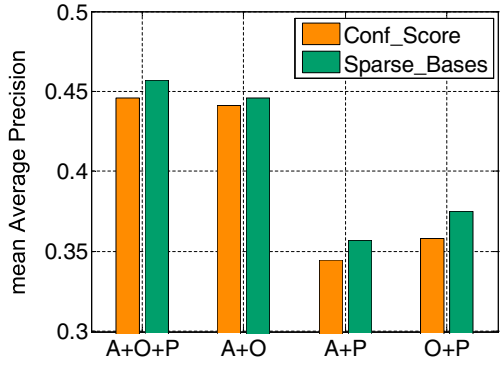

Figure 5. Comparison of the methods by removing the confidence scores obtained from attributes (A), objects (O), and poselets (P) from the feature vector, one at a time. The performance is evaluated using mean Average Precision on the Stanford 40 Actions dataset.

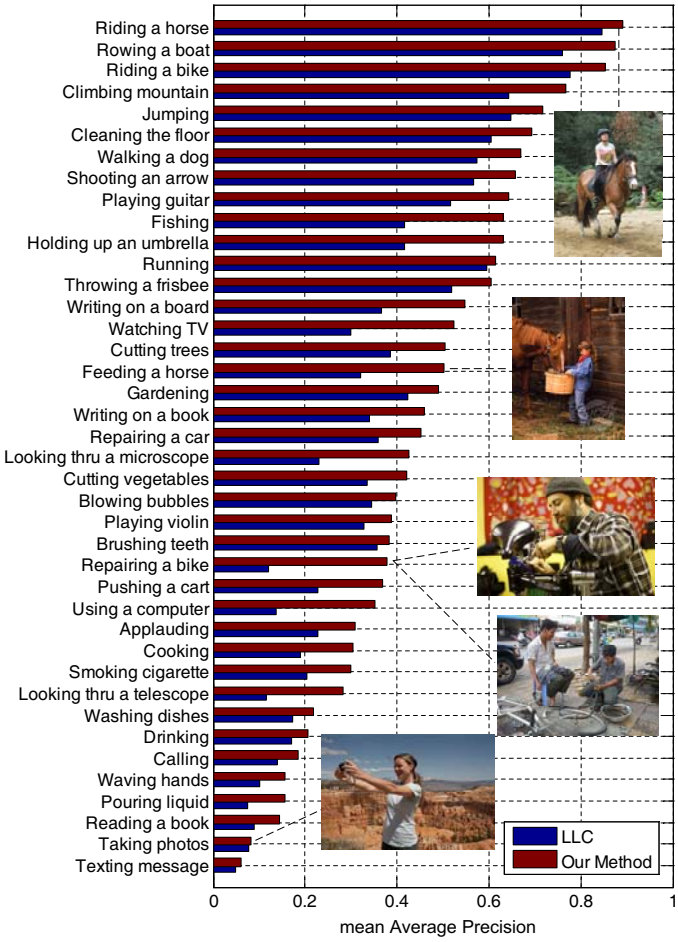

Figure 6. Average precision of our method (Sparse_Bases) on each of the 40 classes of the Stanford 40 Actions dataset. We compare our method with the LLC algorithm.

occluded, and therefore the horse detector is difficult to detect them. From the images of "repairing a bike", we can see that the human pose changes a lot and the bikes are also occluded or disassembled, making them difficult to be recognized by bike detectors. There are some classes on which the recognition performance is very low, e.g. "taking photos". The reason is that the cameras are very small, which makes it difficult to distinguish "taking photos" and the other actions.

6. Discussion

In this work, we use attributes and parts for action recognition. The attributes are verbs related description of human actions, while the parts are composed of objects and poselets. We learn a set of sparse bases of the attributes and parts based image representation, allowing an action image to be reconstructed by a set of sparse coefficients with respect to the bases. Experimental results show that our method achieves state-of-the-art performance on two datasets. One direction of our future work is to use the learned action bases for image tagging, so that we can explore more detailed semantic understanding of human actions in images.

Acknowledgement. L.F-F. is partially supported by an NSF CAREER grant (IIS-0845230), an ONR MURI grant, the DARPA VIRAT program and the DARPA Mind's Eye program. X.J. and L.G. is partially supported by an NSF grant (IIS-101632), an ARO grant (W911NF-10-1-0037) and an ONR MURI grant. B.Y. is partially supported by the SAP Stanford Graduate Fellowship.

References

[1] T. L. Berg, A. C. Berg, and J. Shih. Automatic attribute discovery and characterization from noisy web data. In *ECCV*, 2010. 1, 2, 3

[2] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3D human pose annotations. In *ICCV*, 2009. 1, 2, 3, 5, 11

[3] E. J. Candes and T. Tao. Decoding by linear programming. *IEEE T. Inform. Theory*, 51(12):4203–4215, 2005. 10

[4] V. Delaitre, I. Laptev, and J. Sivic. Recognizing human actions in still images: A study of bag-of-features and part-based representations. In *BMVC*, 2010. 1, 2, 4, 7, 10

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009. 4, 11

[6] C. Desai, D. Ramanan, and C. Fowlkes. Discriminative models for static human-object interactions. In *SMiCV*, 2010. 1

[7] M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2010 (VOC2010) Results. 2, 4, 5

[8] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification. *J. Mach Learn. Res.*, 9:1871–1874, 2008. 10

[9] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *CVPR*, 2009. 1, 2, 3, 4

[10] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE T. Pattern Anal.*, 32(9):1627–1645, 2010. 2, 4, 11

[11] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *ICCV*, 2003. 2

[12] A. Gupta, A. Kembhavi, and L. S. Davis. Observing human-object interactions: Using spatial and functional compatibility for recognition. *IEEE T. Pattern Anal.*, 31(10):1775–1789, 2009. 1, 4

[13] N. Ikizler-Cinbis, R. G. Cinbis, and S. Sclaroff. Learning actions from the web. In *ICCV*, 2009. 1, 4

[14] X. Jiang, Y. Yao, and L. Guibas. Stable identification of cliques with restricted sensing. In *NIPS Workshop on Learning with Orderings*, 2009. 2, 4, 10

[15] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009. 1, 2, 3

[16] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006. 1, 10

[17] H. Lee, A. Battle, R. Raina, and A. Y. Ng. Efficient sparse coding algorithms. In *NIPS*, 2007. 2

[18] L.-J. Li, H. Su, E. P. Xing, and L. Fei-Fei. Object bank: A high-level image representation for scene classification and semantic feature sparsification. In *NIPS*, 2010. 2, 7, 10

[19] J. Liu, B. Kuipers, and S. Savarese. Recognizing human actions by attributes. In *CVPR*, 2011. 2

[20] D. G. Lowe. Distinctive image features from scale-invarient keypoints. *Int. J. Comput. Vision*, 60(2):91–110, 2004. 4, 10

[21] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. *J. Mach. Learn. Res.*, 11:19–60, 2010. 2, 4

[22] S. Maji, L. Bourdev, and J. Malik. Action recognition from a distributed representation of pose and appearance. In *CVPR*, 2011. 1, 2

[23] D. Parikh and K. Grauman. Interactively building a discriminative vocabulary of nameable attributes. In *CVPR*, 2011. 1, 2

[24] A. Prest, C. Schmid, and V. Ferrari. Weakly supervised learning of interactions between humans and objects. Technical report, INRIA, 2010. 1

[25] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. B*, 58(1):267–288, 1996. 2

[26] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In *CVPR*, 2010. 4, 7

[27] W. Yang, Y. Wang, and G. Mori. Recognizing human actions from still images in latent poses. In *CVPR*, 2010. 1

[28] B. Yao and L. Fei-Fei. Grouplet: A structured image representation for recognizing human and object interactions. In *CVPR*, 2010. 1, 2, 4

[29] B. Yao and L. Fei-Fei. Modeling mutual context of object and human pose in human-object interaction activities. In *CVPR*, 2010. 1, 2

[30] B. Yao, A. Khosla, and L. Fei-Fei. Combining randomization and discrimination for fine-grained image categorization. In *CVPR*, 2011. 1

[31] P. Zhao and B. Yu. On model selection consistency of lasso. *J. Mach. Learn. Res.*, 7:2541–2563, 2006. 10

[32] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. B*, 67(2):301–320, 2005. 2, 4

A. The Stanford-40 Action Dataset

Here we give more details of the Stanford 40 Actions dataset collected by us. Our motivation of collecting this dataset is to provide a relatively large scale dataset of daily human actions. The images in the dataset contain large human appearance variation, pose variation, and background clutter. The list of the 40 actions and the number of images in each action is summarized in Table 4. Example images of our dataset are shown in Fig.8.

The images are collected in the following procedure. For each action, we first use some keywords to crawl as many images as we can from Google, Bing, and Flickr. Instead of only using the action name as the keyword, we also consider some other keywords which we believe can help collecting more images of the corresponding action. For example, the query keywords we use for "watching TV" is: "watching television", "man watching TV", "woman watching TV", "family watching TV", and "children watching TV". We can crawl 10,000+ images for each class. Then, we select the desired images from the crawling results in each class. The selection criteria are: (1) the human should be doing the corresponding action; (2) the human's head needs to be visible; (3) the image is not a cartoon image; (4) the image should not be significantly edited (e.g. with many texts on it); (5) the human should not be doing more than one of our 40 actions (e.g. pushing a cart while calling). The next step

Action Name	# imgs	Action Name	# imgs
applauding	284	playing violin	260
blowing bubbles	259	pouring liquid	200
brushing teeth	200	pushing cart	235
cleaning floor	212	reading	245
climbing	295	riding bike	293
cooking	288	riding horse	296
cutting trees	203	rowing boat	185
cutting vegetables	189	running	251
drinking	256	shooting arrow	214
feeding horse	287	smoking	241
fishing	273	taking photos	197
fixing bike	228	texting message	193
fixing car	251	throwing frisby	202
gardening	199	using computer	230
holding umbrella	292	walking dog	293
jumping	295	washing dishes	182
looking thru microscope	191	watching TV	223
looking thru telescope	203	waving hands	210
phoning	259	writing on board	183
playing guitar	289	writing on book	246

Table 4. The Stanford 40 Action dataset: the list of actions and number of images in each action.

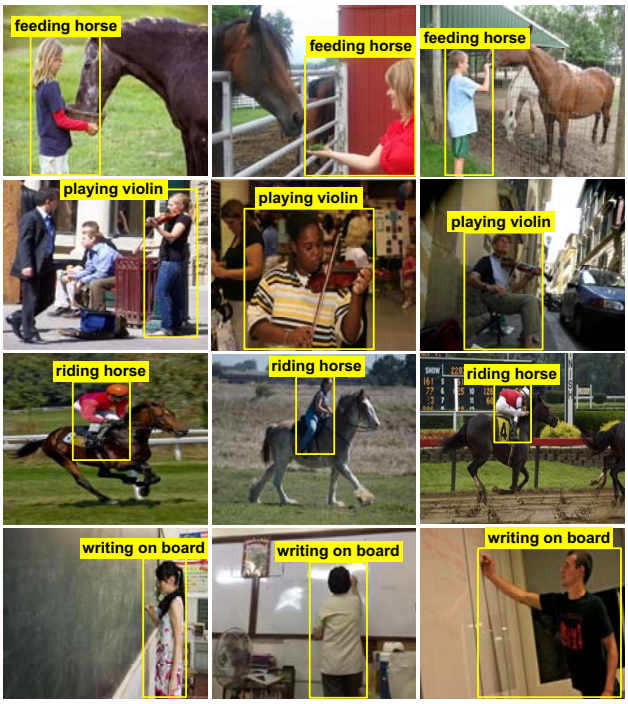

Figure 7. Example images in the Stanford 40 Action dataset.

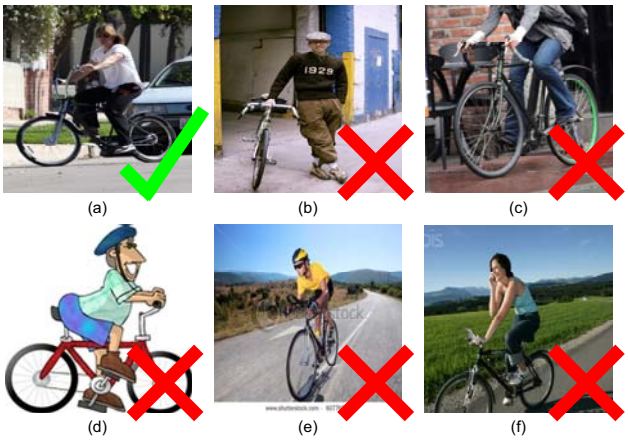

Figure 8. In the case of "riding a bike", we want to collect images such as (a). The other images are not satisfying because: (b) the human is not riding the bike; (c) the human's head is totally outside of the image; (d) it is a cartoon image; (d) it is an advertisement image and texts are placed on the image; (e) the human is riding a bike while making a phone call.

is to de-duplicate the selected images by a simple color histogram matching method. Finally, we check the remaining images and further manually remove some images to guarantee the image diversity within each class.

B. Optimization Guarantees of Dual Sparsity

We used dual sparsity on both action bases and reconstruction coefficients in the basis learning step (Eqn.2 and 4 of the main paper). We now show that this dual sparsity enables the uniqueness of the attributes and parts reconstruction in the testing step (Eqn.3 of the main paper). Uniqueness is important in the Lasso problem especially when we look for interpretable bases for action recognition. Otherwise if the solution for the problem is not unique, one might reconstruct the attributes and parts of an action image from other confusing bases which also optimize our objective but are totally irrelevant to the action in the image.

It has been shown that the ℓ_1-norm minimization problem has a unique sparse solution, if the basis matrix satisfies the so-called *Restricted Isometry Property (RIP)* condition, which requires that every subset of columns in the support of the sparse signal are nearly orthogonal [3]. In [31], the *Irrepresentable Condition* (*IRR*) was proposed for stably recovering a sparse signal $\mathbf{w}^*$ by solving the Lasso problem:

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{a} - \mathbf{\Phi}\mathbf{w}\|_2^2 + \lambda\|\mathbf{w}\|_1. \tag{6}$$

The basis matrix $\mathbf{\Phi}$ satisfies the IRR condition with respect to $S = \{\forall j, w_j^* \neq 0\}$, if $\mathbf{\Phi}_S^T\mathbf{\Phi}_S$ is invertible and

$$\|\mathbf{\Phi}_{S^c}^T\mathbf{\Phi}_S\left(\mathbf{\Phi}_S^T\mathbf{\Phi}_S\right)^{-1}\|_\infty < 1. \tag{7}$$

where $\mathbf{\Phi}_S$ is a sub-matrix of $\mathbf{\Phi}$ with S selecting the columns, $\mathbf{\Phi}_S^T$ is the transpose of $\mathbf{\Phi}_S$, S^c is the complement of S.

The IRR condition is satisfied in some specific situations, such as Gaussian random matrices. But it does not hold for general matrices. However, it has been shown that when the basis matrix $\mathbf{\Phi}$ is sparse, it turns out that IRR still holds in many different situations [14]. Please refer to [14] for further materials explaining the conditions to guarantee the success of solving the Lasso problem. In our problem, we impose sparsity on $\mathbf{\Phi}$ so that a unique sparse solution of $\mathbf{w}$ can be obtained for most of the vectors $\mathbf{a}$.

C. Implementation Details of Our Experiment

In this work, we use attributes, objects, and poselets for action recognition. In Sec.5.1, we have described how we use them for image representation. More details are provided below.

C.1. Attribute Representation

We use 14 attributes for the PASCAL data and 45 attributes for the Stanford 40 data. The list of attributes on the two datasets are:

- Attributes on PASCAL Actions: calling, playing, reading, riding, running, taking, using, walking, cycling, standing, sitting, squatting, lying, and moving.
- Attributes on Stanford 40 Actions: applauding, bending, blowing, brushing, calling, cooking, cutting, cycling, drinking, feeding, fishing, fixing, filling, hanging, holding, jumping, looking through, lying, mopping, playing, poling, pulling, pushing, reading, repairing, riding, rowing, running, shooting, singing, sitting, smoking, speaking, standing, squatting, taking, throwing, typing, using, walking, watching, waving, wearing, withdrawing, and writing.

We train an SVM classifier for each action attribute, where the humans whose actions are described by the attribute are regarded as positive examples, while others are negative examples. Following the approach in [4], each human is described by a "foreground" image (an extension of the bounding box of the human) and a "background" image (the whole image that the human belongs to). We extract SIFT [20] descriptors from the images, and use the Locality-constrained Linear Coding method for feature representation. We use a four-layer spatial pyramid [16] on the foreground and a two-layer pyramid on the background. We use LIBLINEAR [8] for SVM training.

C.2. Object Representation

Whether an image appears in an image is represented by the confidence of object detection scores. We take the object detectors trained from the object bank [18] method. The detectors we use for the two datasets are:

- Objects on PASCAL Actions: beach, bicycle, bicycle built for two, camcorder, camera, cello, cellular telephone, computer, computer keyboard, desktop computer, dial telephone, flute, grass, guitar, keyboard, laptop, monitor, motorcycle, musical instrument, newspaper, notebook, pay phone, piano, skyscraper, telephone, and violin.
- Objects on Stanford 40 Actions: African hunting dog, Eskimo dog, Polaroid camera, beach, beer, beer bottle, beer glass, bicycle, bicycle built for two, blackboard, boat, boathouse, bow, bowl, broom, bulldog, camcorder, camera, car-12982, car-1527, car-1634, car tire, coat, computer, computer keyboard, cup, cuppa, desktop computer, dog, fish, fishing rod, gas pump,

glass, golden retriever, grass, guitar, hand-held computer, handcart, laptop, laundry cart, male horse, motorcycle, mountain bike, mug, newspaper, notebook, optical telescope, passenger car, point-and-shoot camera, radio telescope, sheet, shopping cart, sky, streetcar, television, violin, washbasin, washer, and wheel.

The objects we consider are limited to the ones that have annotated bounding boxes in ImageNet [5]. For instance, "car-12982", "car-1527", and "car-1634" are three different cars in ImageNet. For each object, there is a corresponding deformable part detector [10] in object bank. Each detector consists of two components with six scales each component. Therefore if there is N objects, the dimension of the object feature will be $12N$.

C.3. Pose Representation

We use the 150 pre-trained poselet detectors provided in [2]. This gives us a 150-dimensional pose feature representation on each image.