

Recognition in Activity Space by Connecting Verbs, Nouns, and Human Poses

Paper ID 255

Abstract

While detection and classification of nouns (in most cases objects) has been enjoying great progress in the past ten years, there is relatively little understanding in computer vision of human activities using verbs and the nouns (i.e. objects) associated with them. In order to represent a semantically meaningful space of human activities, we observe the necessity of using visually rich human poses to link the meanings of verbs and nouns. Given training data, we first learn a dictionary of many different poses, which we call “atomic poses”, and construct a joint model for the verbs, nouns and atomic poses. The result is a set of learned connectivity among these three concepts, characterizing the space of different human activities. Using this connectivity graph, we can 1) build meaningful taxonomies of verbs and nouns in the activity space, 2) cluster and categorize different types of activities based on the verb-pose-noun interaction, and 3) recognize different components of the activity (object detection, pose estimation). Our model outperforms the state-of-the-art algorithms in a number of tasks using the HOI Sports [14] and PASCAL action [9] datasets.

1. Introduction

Human activities are often described using verbs or phrases of verb and nouns, such as when the person is “walking” or “playing basketball”. While computer vision research has made great progress in recognizing nouns [9, 11], scaling up to tens of thousands of object classes [23, 5]), recognizing verbs that describe human activities is still a wide open area of research.

Psychologists have shown that infants learn nouns much earlier than verbs [13]. One hypothesis is that most nouns refer to existing entities, often easily accessible to humans. The verbs, on the other hand, tend to describe the concepts that are more abstract and diffuse [12]. In this paper, we aim to jointly model verbs and nouns and explore their relationship, which provides a novel representation of human activities.

An illustration of our ideas is shown in Fig.1. The key insight is that *verbs* and *nouns* are often impoverished linguistic symbols for characterizing human activities. For example, the phrase “serve a volleyball” can render different im-

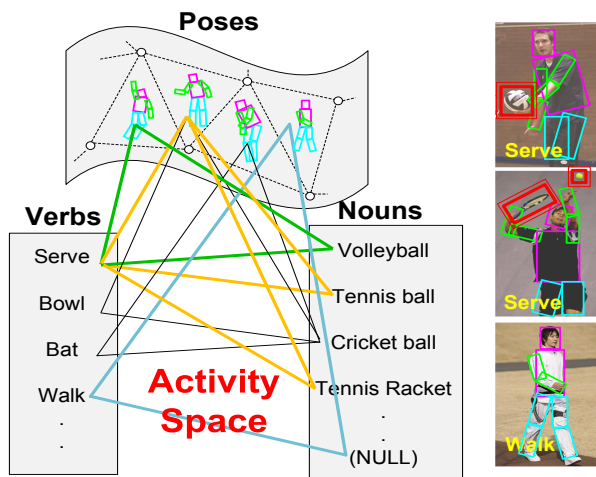


Figure 1. We represent a human action as a triangulation between a verb, a noun, and a human pose. Our method explicitly learns semantically meaningful connections between sets of verbs, nouns, and human poses. We also deal with the cases where multiple nouns or no noun is included.

ages depending on whether the serve is overhand or underhand. There are also linguistic ambiguities between verbs and nouns where one verb corresponds to different nouns and vice versa; for example, one can serve both a volleyball or a tennis ball, as shown in Fig.1. To better characterize human activities we introduce a third dimension that encodes rich visual information: *human poses*, which serve as a visual link between the verbs and the nouns. Possible layouts of body parts are discretized into a set of representative poses that distinguish the ambiguities of verb-noun combinations; for example, “talking on the phone” has a large variety of visual appearances depending on the person’s pose. We obtain a dictionary of spatial layouts of body parts which we call *atomic poses* (see Fig.2 for examples).

Building upon this intuition, we represent human activities as interactions of verbs, nouns, and poses. We observe that connecting these three concepts establishes a semantically meaningful activity space, where recognition of specific components (e.g. objects, or human poses) or clustering of different types of activities are a natural outcome. Our work is related to a handful of recent work on recognizing human-object interaction (HOI) activities in still images [26, 7, 21, 14]. While our work is synergistic, it



(a) Examples of atomic poses obtained from the sports dataset [14].



(b) Examples of atomic poses obtained from the PASCAL dataset [9].

Figure 2. We show four atomic poses on each data set; notice the similarity of human poses in each atomic pose. This figure also shows that similar poses might represent different verbs and convey different semantic meaning. For example, the last atomic pose of (a) corresponds to three different activities (volleyball smashing, cricket bowling, and tennis serving).

does not assume the existence of objects in every activity (e.g. walking) and can accommodate multiple objects (e.g. serving tennis ball with a racket), as shown in Fig.1. Furthermore, we emphasize three goals of this paper that are beyond object detection, pose estimation, and activity classification in the previous HOI work:

- *Constructing* an activity space of verbs, nouns, and human poses, and explicitly learning a semantically meaningful association between them (Sec.3.2.3).
- *Measuring* distance and *building* taxonomies of verbs and nouns¹ that consider both the semantic meaning and the image features in human activities (Sec.3.2.3).
- *Clustering* images in the activity space, where images of similar activities are naturally closer to each other due to the interactions of verbs, poses and objects (Sec.4.3).

The rest of the paper is organized as follows. We introduce related work in Sec.2, and elaborate on our approach for learning the verb-noun-pose associations and for constructing the activity space in Sec.3. In Sec.4 we represent experimental results.

¹We use **nouns** and **objects** interchangeably, if not explicitly stated otherwise since most nouns involved in human activities are objects.

2. Related Work

Activity recognition in still images is mostly treated as an image classification problem, without a detailed understanding of the different components in the image (see [18] for a thorough review). Recently, contextual information has been used to achieve better human activity recognition performance, such as human faces [16], scene background [15], objects, and human poses [14, 26]. Activity recognition is more widely studied in videos, where spatial temporal information is used for identifying classes of motions [2]. However, this is outside of the scope of this paper since our goal is to map the activity space using poses, verbs, and nouns modeled from still images.

There has been work on relating language to pictures, mostly using nouns [8], adjectives [17, 10] and propositions [14]. Little has been done to model verbs and to represent human activity space using verbs and nouns.

In order to effectively model verbs and nouns, our method builds upon state-of-the-art object detection [11, 6], pose estimation [22], and image classification [19] approaches. In particular, our model draws inspiration from the mutual context model [26]. However, we focus on learning a universal representation of verbs, poses and nouns that enables the construction of an activity space and verb taxonomy, in addition to the recognition tasks in [26]. Therefore unlike [26], our learning objective does not optimize for a classification task.

We propose using the atomic poses as an intermedia of the verb-noun interactions. The set of atomic poses can be thought of a dictionary of human poses. We use a semi-supervised method to obtain the atomic poses, where only annotations of human body parts are used. This is in contrast to the unsupervised approach [24] and the fully supervised approach where annotations of both body parts and action classes are used [26]. Our atomic poses are discovered in a way similar to poselets [4, 3, 25], which are detectors for specific body parts. In contrast, the atomic poses carry higher-level information by forming the vocabulary for the whole human body.

3. Modeling the Activity Space by Joint Representation of Verbs, Nouns, and Poses

To represent a human activity as a verb-noun-pose interaction, we construct a three dimensional activity space based on a set of verbs \mathcal{V} (such as “serve” and “ride”), a set of nouns or objects \mathcal{O} (such as “volleyball” and “guitar”), and a set of atomic poses \mathcal{H} that form a dictionary of human body layouts (see Fig.2 for examples of atomic poses and Sec.3.2.1 for how to obtain them). The space is constructed in the way such that two images are close to each other if the activities depicted in them are semantically related (e.g. similar poses, same objects, or related verbs).

In order to construct this space, we first propose a model (Sec.3.1) to characterize the relationships between verbs, nouns, and atomic poses. Our model learns a universal connectivity (Sec.3.2.2) between them, where stronger interactions are depicted by large weights in the connection potential (e.g. “ride” and “bike” are expected to have a strong connection). We then use a probabilistic interpretation of this connectivity to build a verb taxonomy and a noun taxonomy of human activities (Sec.3.2.3). Combining the taxonomies with the set of atomic poses, we build the three dimensional verb-noun-pose space of human activities (Sec.3.2.3).

3.1. Model Details

Given an image I with the annotations of verb $V \in \mathcal{V}$, bounding boxes of nouns $O \in \mathcal{O}$ and body parts in the human pose $H \in \mathcal{H}$, our model learns the connectivity between verbs, nouns, and atomic poses that reflects the strength of the interactions between them. We further make the connectivity conditioned on the image evidence, so that concepts that are harder to recognize will play a less important role in the interaction. Our model is represented as

$$\Psi(V, O, H, I) = \underbrace{\phi_1(V, O, H)}_{\text{v-n-p comp.}} + \underbrace{\phi_2(V, I)}_{\text{verb}} + \underbrace{\phi_3(O, I)}_{\text{noun}} + \underbrace{\phi_4(H, I)}_{\text{pose}} + \underbrace{\phi_5(O, H)}_{\text{spatial relat.}} \quad (1)$$

where ϕ_1 models the compatibility between V , O , and H ; ϕ_{2-4} models the image evidence using the state-of-the-art image classification [19], object detection [11], and pose estimation approaches [1]; ϕ_5 considers the spatial relationship between the objects and body parts. We now go through each term in Eq.1. Please refer to the supplementary document for more details.

Compatibility between verb, noun, and pose. The compatibility between V , O , and H reflects the strength of interaction between them. $\phi_1(V, O, H)$ is parameterized as

$$\phi_1(V, O, H) = \sum_{i=1}^{N_h} \sum_{m=1}^M \sum_{j=1}^{N_o} \sum_{k=1}^{N_v} \mathbf{1}_{(H=h_i)} \cdot \mathbf{1}_{(O^m=o_j)} \cdot \mathbf{1}_{(V=v_k)} \cdot \zeta_{i,j,k} \quad (2)$$

where N_h is the total number of atomic poses (see Sec.3.2.1) and h_i is the i -th atomic pose in \mathcal{H} (similarly for N_o , o_j , N_v , and v_k). M is the number of object bounding boxes within the image, and O^m is the object label of the m -th box.

Modeling verbs. $\phi_2(V, I)$ is parameterized by training a verb classifier based on the image region in the bounding

box of the human. We have

$$\phi_2(V, I) = \sum_{k=1}^{N_v} \mathbf{1}_{(V=v_k)} \cdot \eta_k^T \cdot s(I) \quad (3)$$

where $s(I)$ is the output of an SVM classifier [19].

Modeling nouns. Inspired by [6], we model the nouns (objects) in this image using the object detection score in each object bounding box and the spatial relationships between all these boxes. Denoting the detection score of the m -th bounding box as $g(x^m)$, $\phi_3(O, I)$ is parameterized as

$$\phi_3(O, I) = \sum_{m=1}^M \sum_{j=1}^{N_o} \mathbf{1}_{(O^m=o_j)} \cdot \gamma_j^T \cdot g(x^m) + \sum_{m=1}^M \sum_{m'=1}^M \sum_{j=1}^{N_o} \sum_{j'=1}^{N_o} \mathbf{1}_{(O^m=o_j)} \cdot \mathbf{1}_{(O^{m'}=o_{j'})} \cdot \gamma_{j,j'}^T \cdot b(x^m, x^{m'}) \quad (4)$$

where $b(x^m, x^{m'})$ is a bin function as in [6] that encodes the relative geometric configurations of object bounding boxes x^m and $x^{m'}$ using a grid representation.

Modeling poses. $\phi_4(H, I)$ models the atomic pose that H belongs to and the likelihood of observing this image given the atomic pose. We have

$$\phi_4(H, I) = \sum_{i=1}^{N_h} \sum_{l=1}^L \mathbf{1}_{(H=h_i)} \cdot \left(\alpha_{i,l}^T \cdot p(\mathbf{x}_H^l | \mathbf{x}_{h_i}^l) + \beta_{i,l}^T \cdot f^l(I) \right) \quad (5)$$

where $p(\mathbf{x}_H^l | \mathbf{x}_{h_i}^l)$ is the Gaussian likelihood of observing \mathbf{x}_H^l , the joint of the l -th part in H , given the standard location of the l -th part in atomic pose h_i . $f^l(I)$ is the output of a detector [11] for the l -th body part in this image.

Spatial relationships between objects and body parts. We achieve a better modeling of objects and human body parts by considering their spatial relationships. $\phi_5(H, O)$ is parameterized as

$$\phi_5(H, O) = \sum_{m=1}^M \sum_{i=1}^{N_h} \sum_{j=1}^{N_o} \sum_{l=1}^L \mathbf{1}_{(H=h_i)} \cdot \mathbf{1}_{(O^m=o_j)} \cdot \lambda_{i,j,l}^T \cdot b(x^m, \mathbf{x}_H^l) \quad (6)$$

where $b(x^m, \mathbf{x}_H^l)$ denotes the spatial relationship between the m -th object bounding box and the l -th body part in H . We again use the bin function of [6].

3.2. Constructing the Activity Space

Now that we have introduced the basic model formulation, we show how verbs, nouns and poses can be put together for constructing the activity space. Our goal is to use this representation for discovering taxonomies and patterns of activities, as well as for performing recognition task related to the image contents.

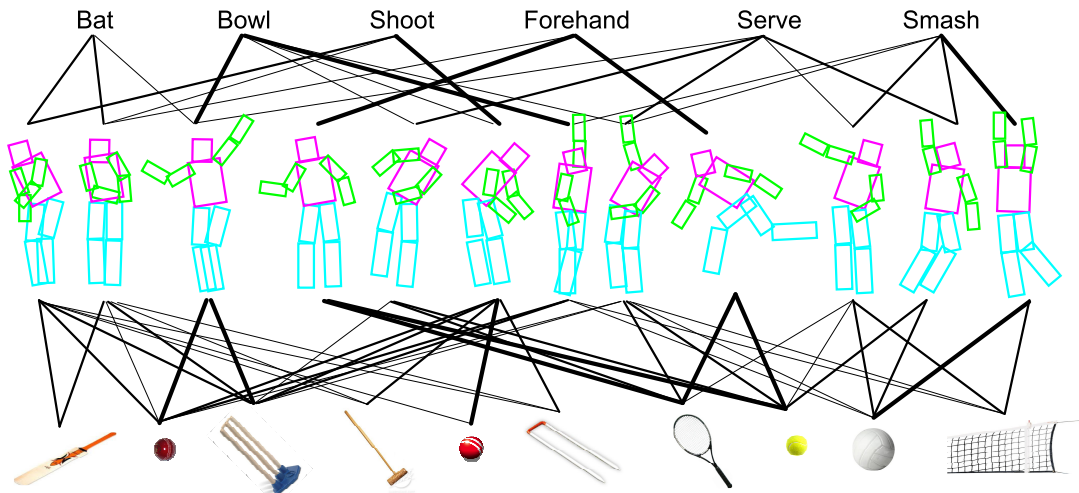


Figure 3. The learned connectivity map of verbs (top row), poses (middle row), and nouns (bottom row) using the sports [14] data set. Thicker lines indicate stronger connections and thinner lines indicate weaker connections. We did not show the connections between verbs and nouns to avoid clutter; we also ignore connections that are too weak.

3.2.1 Obtaining the Atomic Poses

We discuss here in details on how to obtain the atomic poses, a set of the possible layouts of human body parts, using a clustering method. Given the training images, We first align the annotations of each image so that the torsos of all the humans have the same position and size, and constrain the range of variations of both position and angle to $[-1, 1]$. If there is a missing body part due to occlusion, we fill in the annotation with the average annotation values for that particular part. We then use hierarchical clustering with the max linkage measure to obtain a set of clusters. Each cluster represents an atomic pose. Given two images i and j , their distance is measured by

$$d(i, j) = \sum_{l=1}^L \mathbf{w}^T \cdot |\mathbf{x}_i^l - \mathbf{x}_j^l| \quad (7)$$

where \mathbf{x}_i^l denotes the position and orientation of the l -th body part in image i , \mathbf{w} is a weight vector, and L is the number of body parts. The weight for location and orientation are set to 0.15 and 0.1 respectively to account for the fact that empirically the variation range of orientation is larger than that of location. We obtain 12 atomic poses on the sports [14] and the PASCAL dataset [9] respectively. Examples of the obtained atomic poses are shown in Fig.2.

3.2.2 Learning and Analyzing the Connectivity

Our model (Eq.2) is a standard CRF with no hidden variables. We use a belief propagation method [20] with Gaussian priors to learn the model parameters. All object detectors and body part detectors are trained using the deformable parts model [11], and the verb classifier is trained

using the spatial pyramid method. A constant 1 is appended to each feature vector so that the model can learn biases between different classes.

While ϕ_{2-4} estimate the parameters for recognizing verbs, nouns, and human poses individually, ϕ_1 learns the strength of compatibility between them. For example, given the observation of an object, we can use ϕ_1 to predict the verb describing the action and estimate the corresponding pose. Fig.3 visualizes the connectivity that is learned from the sports data set [14]. Each connection is obtained by marginalizing ζ in Eq.2 with respect to the other concept. For instance, the strength of the connection between the verb v_k and noun o_j is estimated by $\sum_{i=1}^{N_h} \exp(\zeta_{i,j,k})$.

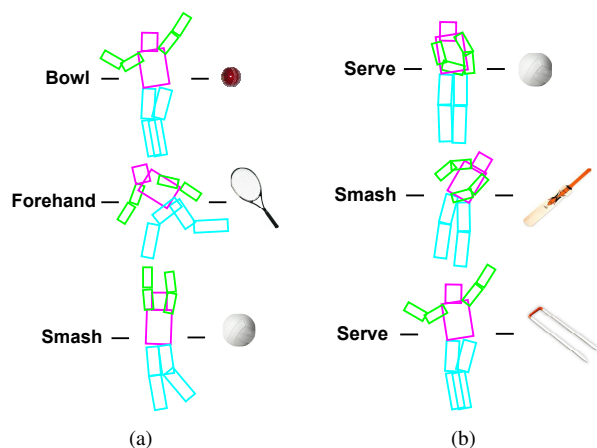


Figure 4. Triplets of verb-pose-noun generated from sampling the learned model (i.e. the connectivity map illustrated in Fig.3): (a) examples sampled from our learned distribution; (b) examples sampled from a uniform distribution.

Fig.3 shows that our method can successfully learn meaningful verb-noun-pose interactions, such as the connection between “forehand” and the fourth atomic pose which is a reasonable gesture for the activity of “tennis forehand”, the noun “volleyball” and the last atomic pose, etc. To further analyze the quality of the connectivity, we take a probabilistic interpretation of our model and use Gibbs sampling to sample two triples of verbs, poses, and nouns. In comparison we also sample from a uniform distribution where all verbs, nouns, and poses are connected with equal weights. Fig.3 shows that we can sample semantically meaningful triples of verbs, nouns, and poses using our learned weights but not the uniform ones.

3.2.3 Constructing the Taxonomies and Activity Space

Our learned verb-noun-pose connectivity offers a way to generate a **verb taxonomy** of human activities. Using a probabilistic interpretation of this connectivity, we represent each verb as a joint distribution of nouns and atomic poses. We can now measure the “distance” between each pair of verbs, from which we build the verb taxonomy. Given a verb v_k , the likelihood of observing pose h_i and noun o_j is estimated as:

$$p(H = h_i, O = o_j | V = v_k) = \frac{1}{Z(v_k)} \exp(\zeta_{i,j,k}) \quad (8)$$

where $Z(v_k) = \sum_{i=1}^{N_h} \sum_{j=1}^{N_o} \exp(\zeta_{i,j,k})$. Then the “distance” between two verbs v_k and $v_{k'}$ is estimated as

$$D_v(v_k, v_{k'}) = D_{KL}(p(H, O | v_k) || p(H, O | v_{k'})) + D_{KL}(p(H, O | v_{k'}) || p(H, O | v_k)) \quad (9)$$

where $D_{KL}(p || p')$ is the K-L divergence between p and p' .

With this distance measure, we use hierarchical clustering with the min linkage measure to build the verb taxonomy. The verb taxonomies that are estimated from the sports dataset [14] and the PASCAL action dataset [9] are shown in Fig.5(a) and Fig.5(c) respectively. Comparing with Fig.5(b) and Fig.5(d), where the taxonomies are built by only using the distribution of atomic poses in each activity, our taxonomies better captures the semantic meaning of different verbs. For example, in Fig.5(a), “forehand” and “serve” are neighbors due to their interaction with the same set of nouns (“tennis racket” and “tennis ball”). They are also close to “smash” because the human pose of “smash” is similar to the pose of “serve”. Similarly, two related verbs “walk” and “run” in the PASCAL dataset whose corresponding activities do not explicitly involve any objects are placed together by our taxonomy in Fig.5(c) but not by the taxonomy relying only on human poses in Fig.5(d).

In order to construct the **activity space**, we also define distance measures and construct taxonomies for nouns and poses in a way that the images that are close together correspond to related nouns or have similar poses. For nouns,

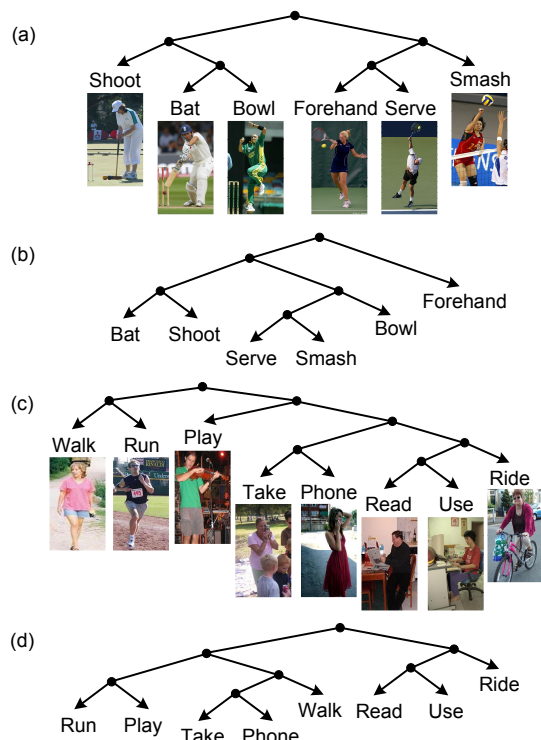


Figure 5. Comparison of verb taxonomies obtained from: (a) our method on the sports dataset [14]; (b) only using pose distribution on the sports dataset; (c) our method on the PASCAL dataset [9]; (d) only using pose distribution on the PASCAL dataset.

we adopt the same way as that in verbs by considering a distribution $p(H = h_i, V = v_k | O = o_j)$ for each object o_j . Considering that human poses lie in a more continuous visual feature space, we use a more direct way to obtain the pose taxonomy by continuing the clustering procedure described in Sec.3.2.1. The distance between two atomic poses are computed as in Eq.7.

By using the taxonomies of verbs, nouns, and atomic poses, we can construct a 3D space of human activities. This space jointly considers the three components that are critical for human activity recognition. In this space, the distance measure for the verbs and the nouns considers both semantic meaning (as captured by their interaction with the other two components) and the underlying visual features (verb classification, human pose estimation, and object detection). An example of the activity space is shown in Fig.6. We give more analysis and show how to apply this space in recognition tasks in Sec.4.3.

4. Recognition in the Activity Space

4.1. Data Sets and Experiment Settings

We carry out experiments on Gupta’s sports dataset [14] and the PASCAL action dataset [9]. Verbs and nouns in

Activity	Verb	Nouns
cricket batting	<i>bat</i>	<i>cricket bat, ball, and stump</i>
cricket bowling	<i>bowl</i>	<i>cricket ball and stump</i>
croquet shot	<i>shoot</i>	<i>croquet mallet, ball, and hoop</i>
tennis forehand	<i>forehand</i>	<i>tennis racket and ball</i>
tennis serve	<i>serve</i>	<i>tennis racket and ball</i>
volleyball smash	<i>smash</i>	<i>volleyball and net</i>

Table 1. Verbs and nouns in each activity class of the sports data [14]. Every verb interacts with more than one object, and some objects correspond to more than one verb. For example, the objects in “forehand” and “serve” are exactly the same.

Activity	Verb	Nouns
phoning	<i>phone</i>	<i>telephone or cell phone</i>
playing instrument	<i>play</i>	<i>violin, guitar, cello, trumpet, flute, saxophone, or piano</i>
reading	<i>read</i>	<i>book or newspaper</i>
riding bike	<i>ride</i>	<i>bike or motorbike</i>
riding horse	<i>ride</i>	<i>horse</i>
running	<i>run</i>	NULL
taking photo	<i>take</i>	<i>camera</i>
using computer	<i>use</i>	<i>keyboard, monitor, or laptop</i>
walking	<i>walk</i>	NULL

Table 2. Verbs and nouns in each activity class of the PASCAL action data [9]. “NULL” means no noun is involved. The same verb might correspond to different activities (“riding bike” and “riding horse”). Also, one verb might interact with different objects in different images, even though they represent the same activity such as playing musical instrument. Some verbs do not correspond to any nouns (e.g. “run” and “walk”).

different activity classes on each data set are summarized in Table 1 and Table 2. We observe that the two data sets cover many different types of activities, such as ones where the human is interacting with multiple objects or with no object at all. Because our approach involves pose estimation, for the PASCAL dataset we only use the images where at least the upper body parts of the human are visible. Some atomic poses on the two data sets are shown in Fig. 2.

We have a pre-processing step that trains an upper-body detector [11] on each data set. The detectors work almost perfectly in our setting, because the image background of the sports data set is relatively clean and there are bounding boxes of the humans available in the PASCAL action dataset. We normalize the images based on the size of the detection boxes so that we do not need to search over scales for pose estimation. We use the pictorial structure code from [22] in our experiments for pose estimation.

4.2. Pose Estimation, Object Detection, and Activity Classification

The model in Eq. 1 jointly models the verbs, the nouns (objects), and the human poses in an image. We use this

model for conventional recognition tasks including pose estimation, object detection, and activity classification. Inference on Eq. 1 gives us the outputs of pose estimation, object detection, and verb classification. Activity classification results on the sports data set can be directly obtained from verb classification, because every verb only corresponds to one activity on this dataset. On the PASCAL dataset, if the verb of an image is “ride”, we further use horse and bike/motorbike detection results to classify this image as “riding bike” or “riding horse”. Given the initial object and body part detection results, and the SVM outputs of verb classification of a new image, we iteratively perform the following three steps for inference until a local maximum of $\Psi(V, O, H, I)$ is reached. We only give a very brief outline of our inference method due to space limitations. Please refer to the supplementary document for more details.

- Updating the layout of human body parts.** From the current inference result, we compute the marginal distribution of the human pose H over all atomic poses. We use this distribution to refine the priors of the joint locations, and then apply the Pictorial Structure (PS) method [1] to update the layout of human body parts using the new priors.
- Updating the nouns.** With the current pose estimation result and the marginal distribution over atomic poses and activity classes, we use greedy forward search [6] to update the object detection results.
- Updating the verb and atomic pose labels.** Based on the current pose estimation and object detection results, we optimize $\Psi(V, O, H, I)$ by enumerating all possible combinations of H and V labels.

Pose Estimation							
Method	H	T	UA	LA	UL	LL	
mutual context [26]	58	66	44 40	27 29	43 39	44 34	
PS baseline [1]	73	71	46 42	37 38	60 65	61 73	
Our method	78	82	56 47	39 41	62 67	65 81	
Object Detection							
Method	DPM [11]		Multi-objects [6]		Our method		
mAP	41.6%		42.9%		48.3%		
Activity Classification							
Method	[14]	[21]	[26]	SPM [19]	Our method		
Accuracy	79%	83%	83%	100%	100%		

Table 3. Pose estimation, object detection, and activity classification results on the sports data. The body part abbreviations are: H - head, T - torso, UA - upper arms, LA - lower arms, UL - upper legs, LL - lower legs. For pose estimation, detection accuracy of each body parts are reported (omitting “%” due to space limitation). For object detection, we use the mean average precision over all the object classes. And for activity classification, average classification accuracy over all activities are reported. We bold the best performance in each experiment.

Pose Estimation						
Method	H	T	UA	LA	UL	LL
PS baseline [1]	49	55	37 35	27 30	44 42	41 44
Our method	56	60	39 41	32 32	46 50	47 46
Object Detection						
Method	DPM [11]		Multi-objects [6]		Our method	
mAP	13.5%		13.7%		16.9%	
Activity Classification						
Method		SPM [19]			Our method	
Accuracy		67.3%			69.2%	

Table 4. Pose estimation, object detection, and activity classification result on the PASCAL data. Please refer to the caption of Table 3 for interpretations of the results.

The results of pose estimation, object detection, and activity classification on the sports and PASCAL data sets are shown in Tables 3 and 4 respectively. Our method outperforms the baselines by considering the interactions between the verbs, nouns, and poses. In Table 3, even the PS baseline of [1] can do better pose estimation than the mutual context model, probably because the body part detectors that we use are more discriminative, and because we use an upper-body detector to avoid searching over different scales. The activity classification results in Table 3 are surprising. Both spatial pyramid matching (SPM) [19] and our method can achieve 100% accuracy. Note that in our method, the verb classification is based on the outputs of an SPM classifier. Finally, we can see that the performance on the sports data set is generally better than that on PASCAL. This is because the resolutions of many images in PASCAL is very low, and many of the human body parts are occluded.

4.3. Image Clustering in the Activity Space

One of our main contributions of the paper is the construction of an activity space by jointly modeling verbs, poses and nouns. Given an activity space, we can perform many new activity recognition tasks besides the conventional activity classification; for example, the distance measures defined for verbs, nouns, and poses (Sec.3.2.3) make it possible to evaluate the distance between every two images considering the semantic information of human activities. With this distance measure, we can discover meaningful activity clusters, perform activity retrieval and activity annotation, etc. In this paper, we offer preliminary results on the first application - image clustering, which has many potential applications in digital image management and image search.

Fig.6 shows the activity space constructed from the sports dataset. (Please refer to the supplementary document for the activity space on the PASCAL dataset.) To perform image clustering, we first denote the inference results of verbs, nouns, and human poses for an image as (v, o, h) .

Given two images (v, o, h) and (v', o', h') , their distance in the activity space is measured by $D_v(v, v') + D_o(o, o') + D_h(h, h')$, where $D_v(v, v')$ and $D_o(o, o')$ are the K-L divergence as in Eq.9, $D_h(h, h')$ is the distance between the layouts of body parts in the two atomic poses as in Eq.7. With this distance measure, an assortment of clustering algorithms can be applied. For simplicity, we choose a simple k-means clustering method. The results are shown in Fig.6. We observe that the clusters obtained accurately capture the semantic meaning of human activities. In most cases images in the same cluster belong to the same activity class. Some of the activities such as cricket bowling correspond to more than one cluster. The reason is that there are large human variations in these images. But within each cluster, the images still have very similar human poses, as well as the same or related verbs and nouns.

5. Discussion and Conclusion

In this paper, we construct a three dimensional activity space by modeling the verb-noun-pose interactions. This project is just the first step to understanding the semantic meanings of human activities in terms of the relationships between verbs, nouns, and human poses, where there are many more interesting things to explore. Using the verb-noun-pose representations of human activities, our future work is to study the activity space on larger-scale problems where there will be more ambiguities between the three concepts, which makes the construction of verb and noun taxonomies and distance measures more challenging.

References

- [1] M. Andriluka, S. Roth, and B. Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *CVPR*, 2009. 3, 6, 7
- [2] A. F. Bobick and J. W. Davis. The representation and recognition of action using temporal templates. *IEEE T. Pattern Anal.*, 23(3):257–267, 2001. 2
- [3] L. Bourdev, S. Maji, T. Brox, and J. Malik. Detecting people using mutually consistent poselet activations. In *ECCV*, 2010. 2
- [4] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3D human pose annotations. In *ICCV*, 2009. 2
- [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. 1
- [6] C. Desai, D. Ramanan, and C. Fowlkes. Discriminative models for multi-class object layout. In *ICCV*, 2009. 2, 3, 6, 7
- [7] C. Desai, D. Ramanan, and C. Fowlkes. Discriminative models for static human-object interactions. In *CVPR Workshop on SMiCV*, 2010. 1
- [8] P. Duygulu, K. Barnard, J. F. G. de Freitas, and D. A. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *ECCV*, 2006. 2
- [9] M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2010 (VOC2010) Results. 1, 2, 4, 5, 6
- [10] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *CVPR*, 2010. 2

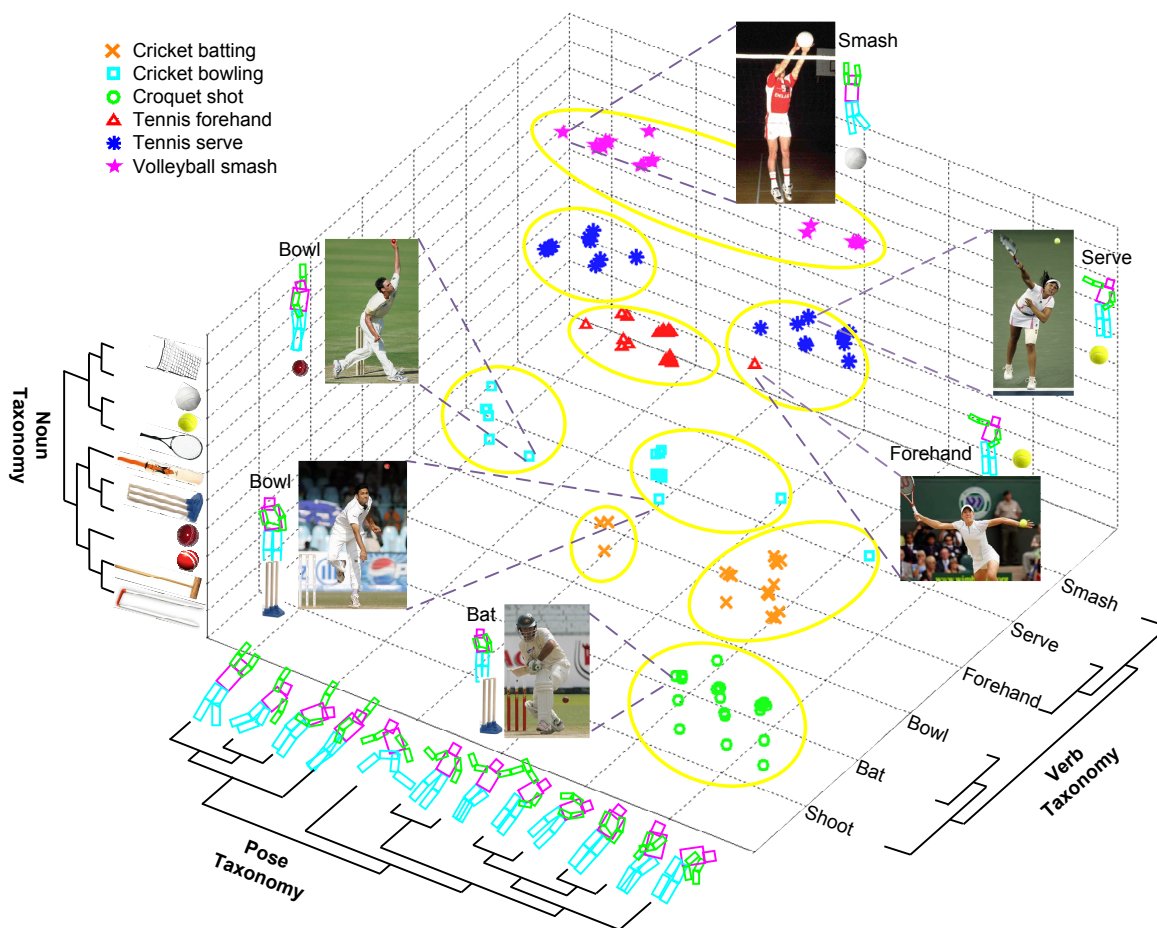


Figure 6. The 3D activity space on the sports dataset [14] and the inference results on the testing images in this space. Each test image is indicated by a colored marker, where the colors and shapes indicate the ground truth labels. The axes correspond to verbs, nouns, and atomic poses, where we also show the corresponding taxonomies for each of these concepts. The coordinates of the test images in this space are obtained by first inferring the verb, noun, and human pose labels for each image (see Sec.4.2), and then projecting the image into the 3D space. We only show the highest scoring object detection result. The distance measure in this space is defined in Sec.3.2.3. Yellow ellipses indicate the k-means clustering boundaries.

- [11] 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. 1, 2, 3, 4, 6, 7
- [12] D. Gentner. Some interesting differences between nouns and verbs. *Cognition and Brain Theory*, 1981. 1
- [13] D. Gentner. *Why Nouns Are Learned Before Verbs: Linguistic Relativity Versus Natural Partitioning*. Language Development, Volume 2: Language Thought and Cluture. Hillsdale, NJ, 1982. 1
- [14] 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), 2009. 1, 2, 4, 5, 6, 8
- [15] N. Ikizler-Cinbis and S. Sclaroff. Object, scene and actions: Combining multiple features for human action recognition. In *ECCV*, 2010. 2
- [16] L. Jie, B. Caputo, and V. Ferrari. Who’s doing what: Joint modeling of names and verbs for simultaneous face and pose annotation. In *NIPS*, 2010. 2
- [17] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009. 2
- [18] I. Laptev and G. Mori. Statistical and structural recognition of human actions. Tutorial on Human Action Recognition at ECCV 2010. 2
- [19] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006. 2, 3, 6, 7
- [20] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference* (2nd ed.). Morgan Kaufmann, 1988. 4
- [21] A. Prest, C. Schmid, and V. Ferrari. Weakly supervised learning of interactions between humans and objects. Technical report, INRIA, 2010. 1, 6
- [22] B. Sapp, C. Jordan, and B. Taskar. Adaptive pose priors for pictorial structures. In *CVPR*, 2010. 2, 6
- [23] A. Torralba, R. Fergus, and W. T. Freeman. 80 million tiny images: A large dataset for non-parametric object and scene recognition. *IEEE T. Pattern Anal.*, 30(11):1958–1970, 2008. 1
- [24] Y. Wang, H. Jiang, M. S. Drew, Z.-N. Li, and G. Mori. Unsupervised discovery of action classes. In *CVPR*, 2006. 2
- [25] W. Yang, Y. Wang, and G. Mori. Recognizing human actions from still images with latent poses. In *CVPR*, 2010. 2
- [26] B. Yao and L. Fei-Fei. Modeling mutual context of object and human pose in human-object interaction activities. In *CVPR*, 2010. 1, 2, 6