Recognizing Human-Object Interactions in Still Images by Modeling the Mutual Context of Objects and Human Poses

Bangpeng Yao, *Member, IEEE,* and Li Fei-Fei, *Member, IEEE*

Abstract—Detecting objects in cluttered scenes and estimating articulated human body parts from 2D images are two challenging problems in computer vision. The difficulty is particularly pronounced in activities involving human-object interactions (e.g. playing tennis), where the relevant objects tend to be small or only partially visible, and the human body parts are often self-occluded. We observe, however, that objects and human poses can serve as mutual context to each other – recognizing one facilitates the recognition of the other. In this paper we propose a *mutual context model* to jointly model objects and human poses in human-object interaction activities. In our approach, object detection provides a strong prior for better human pose estimation, while human pose estimation improves the accuracy of detecting the objects that interact with the human. On a six-class sports dataset and a 24-class people interacting with musical instruments dataset, we show that our mutual context model outperforms state-of-the-art in detecting very difficult objects and estimating human poses, as well as classifying human-object interaction activities.

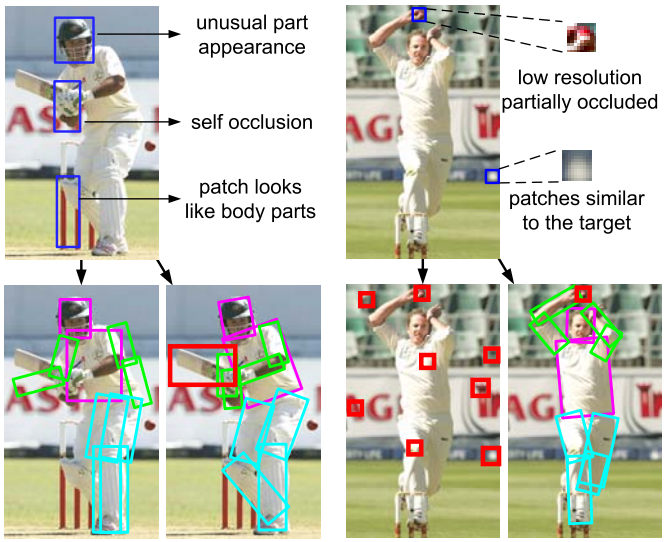
Index Terms—Mutual Context, Action Recognition, Human Pose Estimation, Object Detection, Conditional Random Field

1 Introduction

Using context to aid visual recognition is recently receiving more and more attention. Psychology experiments show that context plays an important role in recognition in the human visual system [1], [2]. In computer vision, context has been used in problems such as object detection and recognition [3], [4], [5], scene recognition [6], action classification [7], and image segmentation [8]. While the idea of using context is clearly a good one, a curious observation shows that most of the context information has contributed relatively little to boost performances in recognition tasks. In the recent Pascal VOC challenge [9], the difference between context based methods and sliding window based methods for object detection (e.g. detecting bicycles) is only within a small margin of $3-4\%$ [10], [11].

One reason to account for the relatively small margin is, in our opinion, the lack of strong context. While it is nice to detect cars in the context of roads, powerful car detectors [12] can nevertheless detect cars with high accuracy no matter whether the cars are on the road or not. Indeed, for the human visual system, detecting visual abnormality out of context is crucial for survival and social activities (e.g. detecting a cat in the fridge, or an unattended bag in the airport) [13].

So is context oversold? Our answer is 'no'. Many important visual recognition tasks rely critically on context. One such scenario is the problem of human pose estimation and object detection in human-object interaction (HOI) activities [14], [15]. As shown in Fig.1, the two difficult tasks can benefit greatly from serving as context for each other. Without knowing that the human is making a defensive shot with the cricket bat, it is not easy to accurately estimate the player's pose (Fig.1(a)); similarly, without seeing the player's pose, it is difficult



(a) Human pose estimation (b) Object (ball) detection

Fig. 1. Objects and human poses can serve as mutual context to facilitate the recognition of each other. In (a), the human pose is better estimated by seeing the cricket bat, which provides a strong prior for the pose of the human. In (b), the cricket ball is detected with the help of understanding the human pose of bowling the ball.

• *B. Yao and L. Fei-Fei are with the Computer Science Department, Stanford University, Stanford, CA 94305.*
E-mail: {bangpeng,feifeili}@cs.stanford.edu

to detect the small ball in the player's hand, which is nearly invisible even to the human eye (Fig.1(b)).

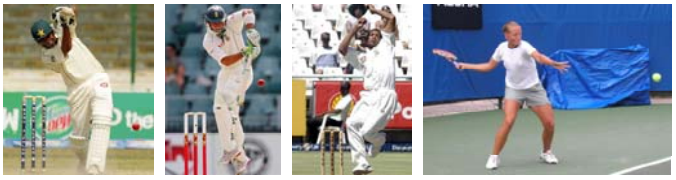
In this paper, we propose to model the *mutual context* between objects and human poses in HOI activities so that each can facilitate the recognition of the other. Specifically, two contextual information are considered in our mutual context model. The *co-occurrence context* models the co-occurrence statistics between objects and specific types of human poses within each activity. The types of human poses, termed as "atomic poses" [16] (shown in Fig.6), can be thought of as a dictionary of human poses, where the human poses represented by the same atomic pose correspond to similar configurations of body parts. We also consider the *spatial context*, which models the spatial relationship between objects and different human body parts. We show that our algorithm significantly improves the performance of both object detection and human pose estimation on a six-class sports dataset [14] and a 24-class people interacting with musical instruments (PPMI) dataset [15]. Furthermore, putting object detection and pose estimation together, our method also achieves higher accuracy in classifying HOI activities.

Modeling the mutual context of objects and human poses has its theoretical basis in psychology. In [17] and [18], it has been shown that humans have a better perception of human gestures when the objects are presented, and vice versa. In [19], the authors confirmed the spatial and functional relationships of objects and human poses in HOI activities. In our work, we explicitly model these relationships so that the recognition of objects and human poses can mutually benefit from each other. This makes our method significantly different from most previous activity recognition approaches, where activity recognition is treated as a pure image or video classification problem [20], [21], [22], [23], without detailed analysis of the objects and human poses that are involved in these activities.

The rest of this paper is organized as follows. Sec.2 describes related work. Details of our model, as well as model learning and inference are elaborated in Sec.3, 4, and 5 respectively. Experimental results are given in Sec.6. Sec.7 concludes the paper.

2 Related Work

Human pose estimation and object detection have been studied in computer vision for many years. Most of the pose estimation approaches model the human body parts in a tree structure and use the pictorial structure method [24], [25] for efficient inference. The pictorial structure approach and its derivations [26], [27], [28], [29] work very well on the images with clean backgrounds and have improved the pose estimation performance in complex scenes such as TV shows. In order to capture more complex human body articulations, some non-tree models have also been proposed [30], [31]. More recently, a real time human pose estimation system has been built by applying the random forest [32] method to depth images [33]. Nevertheless, human pose estimation on 2D images remains a challenging problem, especially when the human body parts are highly articulated and occluded.



(a) The relevant objects that interact with the human may be very small, partially occluded, or tilted to an unusual angle.

(b) Human poses of the same activity might be inconsistent in different images due to different camera angles (the left two images) or the gesture that the human interacts with the object (the right two images).

Fig. 2. Challenges of object detection and human pose estimation in HOI activities.

Sliding window is one of the most successful strategies for object detection. Some techniques have been proposed to avoid exhaustively searching the image [34], [35], which makes the algorithm more efficient. While the most popular detectors are still based on sliding windows [34], [36], more recent work has tried to integrate context to obtain better performance [3], [4], [5]. However, in most of the works the performance is improved by a relatively small margin.

It is out of the scope of this paper to develop an object detection or human pose estimation method that generally applies to all situations. Instead, we focus on the role of context in these problems. Our work is inspired by a number of previous works that have used context in vision tasks [6], [37], [8], [3], [4], [5], [7]. In most of these work, one type of scene information serves as contextual facilitation to a main recognition problem. For example, ground planes and horizons can help to refine pedestrian detections. Specifically, while object context has been widely used to help the recognition of the other objects [3], [10], people have shown that object context can also improve the performance of human pose estimation [38], [39]. Meanwhile, human poses have also been treated as context for many tasks such as motion capture [40] and inferring surface contact (such as joint torques and gravity) [41].

Other than simply treating one task as the main recognition problem and the other one as the contextual facilitation, in this work we explore the *mutual context* between two seemingly unrelated problems - object detection and human pose estimation. Our approach allows the two tasks serve as context for each other, so that the recognition performance of both tasks are

improved. We study the two problems in the activities of human-object interactions in still images, where the *mutual context* plays key roles for understanding the interactions between humans and objects.

Recognizing human activities in still images is a new problem yet received much attention in recent years [14], [15], [42], [43], [44], [45], [46]. While many works treat the task as an image classification problem, more and more people have tried to obtain a detailed understanding of the humans and the objects as well as their interactions. In [47], action recognition is carried out after recognizing the human faces and gestures. In [48], the role of human poses and objects in human activities are analyzed. In [49], the authors propose to directly learn the interactions between humans and objects (or between objects and objects) in a discriminative way, termed as "visual phrases". Our work takes a further step by explicitly modeling the human poses and objects as well as their mutual contexts in HOI activities. Furthermore, we test the performance of object detection, human pose estimation and activity classification in different domains of human activities, including people doing sports [14] and interacting with musical instruments [15].

A preliminary version of our paper was described in [50] and extended in [16]. The model described in this paper is based on [16], which differs from [50] in the following ways. (1) By introducing a set of "atomic poses", we learn an overall relationship between different activities, objects, and human poses, rather than modeling the human-object interactions for each activity separately as in [50]. (2) Instead of limiting to the one human and one object interaction as in [50], the model presented in this paper can deal with the situations where the human interacts with any number of objects (e.g. people interacting with tennis ball and tennis racket in "playing tennis"). (3) The new model incorporates a discriminative action classification component and uses the state-of-the-art object and body part detectors [36], which further improves the recognition performance. Furthermore, in this paper we further test the performance of our method on a people interacting with musical instrument dataset, which is relatively large scale and involves different interactions between the human and the same object (e.g. playing violin versus holding a violin but not playing).

3 The Mutual Context Model

Given an HOI activity, our goal is to estimate the human pose and detect the objects that the human interacts with (shown in Fig.1). Fig.2 illustrates that both tasks are challenging. On one hand, the relevant objects are often small, partially occluded, or tilted to an unusual angle by the human. The human poses, on the other hand, are usually highly articulated where many body parts are self-occluded. Furthermore, even in the same activity, the configurations of body parts might have large variations in different images due to different human gestures or shooting angles of the cameras.

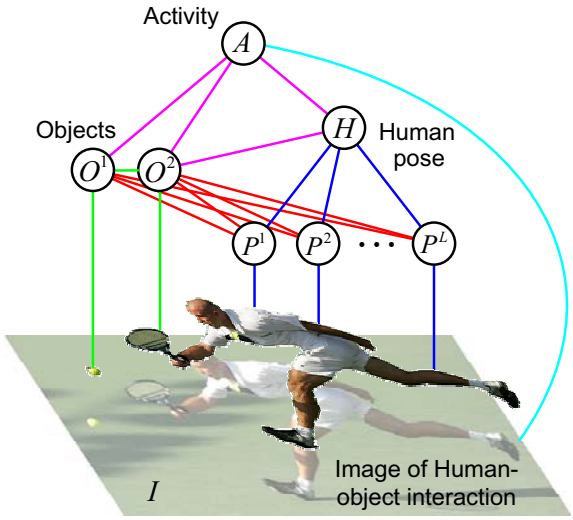


Fig. 3. A graphical illustration of our model. A denotes an HOI activity class, H the atomic human pose class, P a body part, and O the object. We have two O's here because there are two objects in the image (tennis racket and tennis ball). The edges are denoted by different colors as they correspond to different components of our model (detailed in Sec.3.1).

Here we propose a novel model to exploit the mutual context of human poses and objects in one coherent framework, where object detection and human pose estimation can benefit from each other. Our model deals with the situations where the human interacts with any number of objects.

3.1 The Model Representation

A graphical illustration of our model is shown in Fig.3. Our model can be thought of as a conditional random field [51]. For an image I of a human-object interaction activity, our approach jointly models the overall activity class A, the objects $O = \{O^1, \cdots, O^M\}$ interacting with the human, and the human pose H. M is the number of object bounding boxes in the image ($M = 2$ in Fig.3), and O^m is the class label of the m-th box. H indicates the atomic pose (Sec.4.1) label that the human pose belongs to. As shown in Fig.6, the human poses described by the same atomic pose have similar layouts of human body parts. The overall human pose is further decomposed into the spatial layout of some body parts (e.g. head, torso, upper-left arm, lower-right leg, etc.), denoted by $P^1, \cdots, P^L$. Here we assume that the set of atomic poses are given. How to obtain the atomic poses will be introduced in Sec.4.1.

In our mutual context model, the activity classes, objects and human poses all contribute to the recognition and detection of each other. We also make this modeling conditioned on the visual features of the corresponding image regions, so that the components that are harder to recognize play less important roles. Putting everything

together, our model is represented as

$$\Psi(A,O,H,I) = \underbrace{\phi_1(A,O,H)}_{\text{co-occurrence context}} + \underbrace{\phi_2(O,H)}_{\text{spatial context}} + \underbrace{\phi_3(O,I)}_{\text{modeling objects}} + \underbrace{\phi_4(H,I)}_{\text{modeling human pose}} + \underbrace{\phi_5(A,I)}_{\text{modeling activity}} \quad (1)$$

where ϕ_1 models the co-occurrence compatibility between A, O, and H (magenta edges in Fig.3); ϕ_2 considers the spatial relationship between objects and human body parts (red edges in Fig.3); ϕ_{3-5} models the image evidence based on state-of-the-art object detection, human pose estimation, and activity classification approaches (green, blue, and cyan edges in Fig.3, respectively). We now enumerate the potentials of this model:

Co-occurrence context. $\phi_1(A,O,H)$ models the compatibility between the class labels of A, O, and H in terms of co-occurrence frequency. For example, the objects "tennis ball" and "tennis racket" always appear in the activity of "tennis serve", and people usually serve tennis in some specific poses. $\phi_1(A,O,H)$ is parameterized as

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

where N_h is the the total number of atomic poses (see Sec.4.1) and h_i represents the i-th atomic pose; similarly for N_o and o_j, as well as N_a and a_k. $\mathbf{1}_{(\cdot)}$ is an indicator function, e.g. $\mathbf{1}_{(H=h_i)} = 1$ if H equals h_i, otherwise 0. $\zeta_{i,j,k}$ represents the strength of the co-occurrence interaction between h_i, o_j, and a_k: the larger $\zeta_{i,j,k}$ is, the more likely for h_i, o_j, and a_k to co-occur.

Spatial context. The other context we consider is the spatial relationship between objects and different body parts of the human. As shown in Fig.4, in HOI activities, the human pose and object category usually provide a strong and reliable prior for the location of the object with respect to human body parts. We therefore model $\phi_2(O,H)$ by considering each pair of atomic pose and object category. $\phi_2(O,H)$ is parameterized as

$$\phi_2(H,O) = \sum_{m=1}^{M}\sum_{i=1}^{N_h}\sum_{j=1}^{N_o}\sum_{l=1}^{L} \mathbf{1}_{(H=h_i)} \cdot \mathbf{1}_{(O^m=o_j)} \cdot \lambda_{i,j,l}^T \cdot b(\mathbf{x}_I^l, O^m) \quad (3)$$

where $\mathbf{x}_I^l$ is the location of the center of the human's l-th body part in image I, $b(\mathbf{x}_I^l, O^m)$ denotes the spatial relationship between $\mathbf{x}_I^l$ and the m-th object bounding box, and $\lambda_{i,j,l}$ encodes the set of weights for this relationship when the object class of O^m is o_j. We use a binary feature similar to that in [10] to represent $b(\mathbf{x}_I^l, O^m)$. As shown in Fig.5, the relative location of the center of object O^m with respect to $\mathbf{x}_I^l$ is discretized to a set of disjoint regions. Then $b(\mathbf{x}_I^l, O^m)$ is a sparse binary vector with only one 1 for the element that corresponds to the relative location of O^m with respect to $\mathbf{x}_I^l$.

Fig. 4. In HOI activities, the human pose and object category usually provide a strong and reliable prior for the location of the object with respect to human body parts.

Modeling objects. Inspired by Desai et al [10], we model objects in the image using object detection scores in all the object bounding boxes and the spatial relationship between these boxes. Denoting the vector of scores of detecting all the objects in the m-th box as $g(O^m)$, $\phi_3(O,I)$ is parameterized as

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

where γ_j is the weight for the detection scores corresponding to object o_j. $\gamma_{j,j'}$ encodes the set of weights for the geometric configurations between o_j and $o_{j'}$. $b(O^m, O^{m'})$ is a binary feature vector that models the spatial relationship between the m-th and m'-th bounding boxes. We use the same approach as in Fig.5 to obtain $b(O^m, O^{m'})$. Note that in different images, the number of objects (the value of M) can be different.

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

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

where $\alpha_{i,l}$ and $\beta_{i,l}$ are the weights for the location and appearance of the l-th body part in atomic pose h_i. $p(\mathbf{x}_I^l|\mathbf{x}_{h_i}^l)$ is the Gaussian likelihood of observing $\mathbf{x}_I^l$, the joint of the l-th body part in image I, given the standard joint location of the l-th body part in atomic pose h_i. The joints of all the body parts are defined in the same way

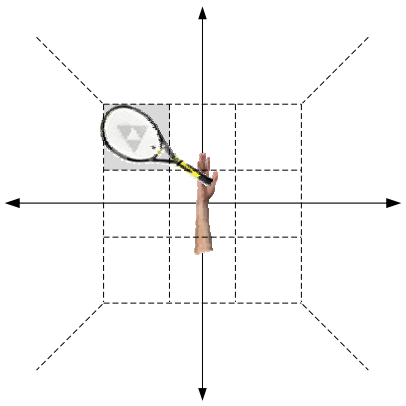


Fig. 5. Visualization of the binary feature $b(\mathbf{x}_I^l, O^m)$. We first divide the space into 13 disjoint regions based on a coordinate frame defined by $\mathbf{x}_I^l$. The image area closer to $\mathbf{x}_I^l$ is divided with a finer resolution. Then $b(\mathbf{x}_I^l, O^m)$ is a 13 dimensional binary vector, with a 1 in the region that the center of O^m falls in (filled by gray color).

as in [24]. $f^l(I)$ is the output for detecting the l-th body part in its current location in this image.

Modeling activities. $\phi_5(A, I)$ takes the global image as features and train an activity classifier to model the HOI activity. It serves as the activity (or scene) context to understand the interactions between humans and objects. We have

$$\phi_5(A, I) = \sum_{k=1}^{N_a} \mathbf{1}_{(A=a_k)} \cdot \eta_k^T \cdot s(I) \tag{6}$$

where $s(I)$ is an N_a-dimensional output of a one-vs-all discriminative classifier. η_k is the set of feature weights corresponding to activity a_k.

3.2 Properties of The Model

Central to our model formulation is the hypothesis that both human pose estimation and object detection can benefit from each other in HOI activities. We highlight here some important properties of our model.

Co-occurrence context for the activity class, object, and human pose. Given the presence of a tennis racket, the human pose is more likely to be playing tennis instead of playing croquet. This is to say, co-occurrence information can be beneficial for jointly modeling the object, the human pose, and the activity class.

Spatial context between objects and body parts. Without knowing the location of the arm, it is difficult to spot the location of the tennis racket in tennis serving. Without seeing the croquet mallet, the heavily occluded arms and legs can become too obscured for robust pose estimation. In HOI activities, different atomic poses imply that the object is handled by the human in different manners, which are modeled by $\phi_2(O, H)$. The modeling is conditioned on the image features, so that we can pay less attention to the objects or human body parts whose corresponding detectors are unreliable.

Flexible to extend to larger scale datasets and other activities. Our model jointly models all the objects and atomic poses in all the HOI activities. Comparing to the original method [50] where the objects and human poses in each HOI activity are modeled separately, our model is easier to extend to larger scale datasets and other activities. Having more activities does not necessarily introduce more atomic poses or objects in our model representation.

Relations with the other models. Our model has drawn inspirations from a number of previous work, such as modeling spatial layout of different image parts [24], [25], [26], using agreement of different image components [3], using multiple models to describe the same concept (human pose in our problem) [52], and discriminative training [26]. Our model integrates all the properties in one coherent framework to perform two seemingly unrelated tasks, human pose estimation and object detection, to the benefit of each other.

4 Model Learning

This section describes our approach for learning different aspects of our model. We first introduce our method of obtaining the atomic poses in Sec.4.1. We then give details of how to train the object and body part detectors, as well as activity classifiers in Sec.4.2. We show how we estimate the model parameters in Sec.4.3.

4.1 Obtaining Atomic Poses

Fig.6 shows three atomic poses. The atomic poses can be thought of as a dictionary of human poses, where the human poses described by the same atomic pose have similar layouts of human body parts. In this paper, the atomic poses play an important role for modeling human and object interactions. Because humans usually manipulate an object in some specific gestures, each HOI activity usually corresponds to some specific atomic poses. Furthermore, human pose estimation can be made much easier if we know which atomic pose that the image corresponds to and hence have a strong and reliable prior of the layout of body parts, as compared to relying on a single model for all the images.

Given a set of training images of HOI activities, we obtain the atomic poses by clustering the configurations of human body parts. We denote the annotation of the human body parts in an image I as $\{\mathbf{x}^1, \cdots, \mathbf{x}^L\}$, where L is the number of body parts and each $\mathbf{x}^l$ is a vector indicating the position and orientation of the l-th body part. We first align the annotations so that the torsos in all the images have the same position and size, and normalize the range of variations of both position and orientation to $[-1, 1]$. If there is a missing body part due to occlusion in an image, we use the annotations of visible body parts to find this image's nearest neighbor, which is used to fill in the annotation of the missing body part. We then use hierarchical clustering with the maximum linkage measure to obtain a set of clusters. Each cluster represents an atomic pose. Given two images I_i and I_j, their distance is measured by $\sum_{l=1}^{L} \mathbf{w}^T |\mathbf{x}_i^l - \mathbf{x}_j^l|$,

Fig. 6. Examples of three atomic poses. Each row shows images of an atomic pose, and the left-most image outlines the representative body parts layout in this atomic pose. Notice the similarity of human poses in each atomic pose. This figure also shows that similar poses might represent different activities. For example, the last atomic pose corresponds to three different activities.

where $\mathbf{w}$ is a weight vector (0.15 and 0.1 for location and orientation components respectively).

Our method for obtaining atomic poses is based on the annotations of human body parts. Therefore it is a "weakly-supervised" approach for clustering human poses, as compared to the previous work where no annotation is used [53] or the clustering is performed within each activity class separately [50], [43]. Compared to the unsupervised approach, the human poses described by each atomic pose are tightly clustered in terms of configurations of body parts, which makes it possible to use the atomic poses for better pose estimation. Compared to the clusters obtained within each activity class separately, our atomic poses are shared by all the activities and are therefore easier to extend to more activity classes.

Our atomic poses are discovered is a way similar to that of poselets [54]. However, while poselets are local detectors for specific body parts, the atomic poses are a dictionary of the overall humans poses. They provide a strong prior of the configuration of human body parts, and therefore can be directly used for better human pose estimation. Fig.7 illustrates all the atomic poses that are obtained from the sports dataset [14]. Fig.10 shows the distribution of the images of some HOI activities over the set of atomic poses obtained from the PPMI dataset [15]. For each atomic pose, we estimate the spatial relationship between different body parts as in the pictorial structure model [24]. These relationships will be used in the inference stage (Sec.5).

4.2 Training Detectors and Classifiers

Our mutual context model is based on a set of object detectors and human body part detectors, as well as an overall activity classifier. In Eq.4, $g(O^m)$ is the score vector of detecting all the objects in the m-th object bounding box. In Eq.5, $f^l(I)$ is the score of detecting the l-th body part in location $\mathbf{x}_I^l$. We train a detector for each object and each human body part using the deformable part model [36]. The deformable part model is a mixture of some discriminatively trained latent SVM classifiers based on the histogram of gradient [55] image features. In our approach, each human body part detector contains one mixture component, while each object detector contains two mixture components unless the aspect ratio of the object does not change at all (e.g. balls). $g(O^m)$ and $f^l(I)$ is the value of the detection score divided by the threshold of the corresponding detector.

The activity classier is trained by using the spatial pyramid matching (SPM) method [56]. We extract SIFT features [57] and apply the histogram intersection kernel on a three layer image pyramid. In Eq.6, $s(I)$ is a N_a-dimensional confidence scores obtained from an SVM classifier, where N_a is the number of activity classes.

4.3 Estimating Model Parameters

In the training stage, we assign each human pose to its closest atomic pose. Given the annotations of human body parts and object bounding boxes, we apply the object and human body part detectors to the corresponding image regions to get detection scores. Therefore our model (Eq.1) is a standard conditional random field with no hidden variables. We use a maximum likelihood approach with zero-mean Gaussian priors to estimate the model parameters $\{\zeta, \lambda, \gamma, \beta, \alpha\}$. Fig.7 and Fig.10 visualizes some of our learning results.

5 Model Inference

Given a new image, inference on Eq.1 gives us the results of activity classification, object detection, and human pose estimation. We initialize the model inference with the SPM action classification results [56], object bounding boxes obtained from independent object detectors [36], as well as initial pose estimation results from a pictorial structure model [28] obtained from all training images, regardless of the belongingness of human poses to different atomic poses. To reduce false negatives in object detection, we keep the detection bounding boxes if the scores are larger than 0.9 of the detection threshold after non-maximum suppression. We then iteratively perform the following steps.

Updating the layout of human body parts. From the current inference result, we compute the marginal distribution of the human pose over all atomic poses: $\{p(H = h_i)\}_{i=1}^{N_h}$. From this distribution, we refine the prior of the joint location of each body part l in this image using a mixture of Gaussians $\sum_{i=1}^{N_h}[p(H = h_i) \cdot \mathcal{N}(\mathbf{x}_{h_i}^l)]$,

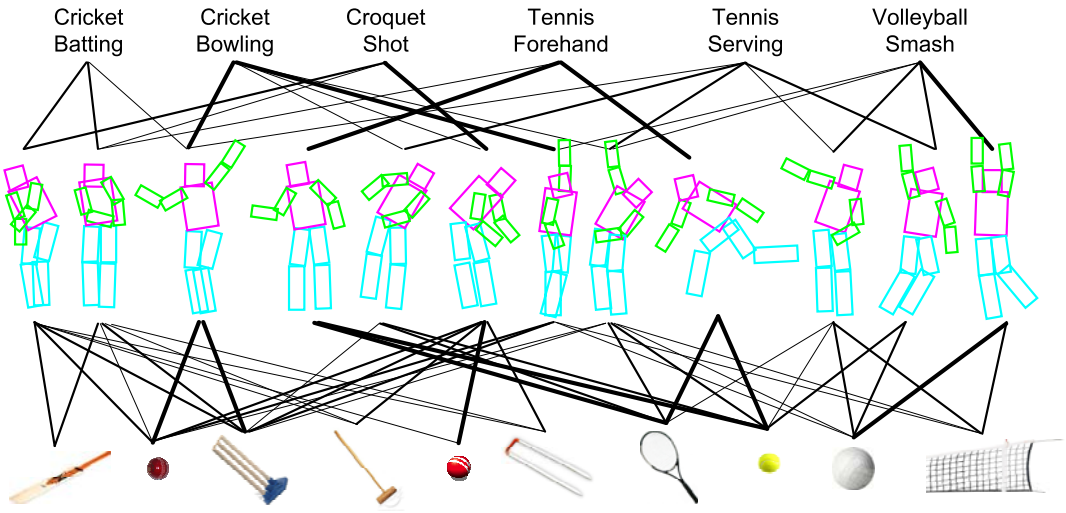


Fig. 7. The learned strength of connectivity between activities, human poses, and objects on the sports dataset [14]. Thicker lines indicate stronger connections. We did not show the connections between activities and objects because they are tricky (e.g. "tennis serving" connects with "tennis ball" and "tennis racket"). We also ignore connections that are very weak.

where $\mathcal{N}(\mathbf{x}^l_{h_i})$ is the prior distribution for body part l in the i-th atomic pose estimated in Sec.4.1. Because the pictorial structure inference can be very efficient if the part dependencies are Gaussians, we further use a Gaussian distribution to approximate each mixture of Gaussians. Then we use pictorial structure with these new Gaussian distributions to update the pose estimation results.

Updating object detection results. With the current pose estimation result as well as the marginal distribution of atomic poses and activity classes, we use a greedy forward search method [10] to update the object detection results. We use (m, j) to denote the score of assigning the m-th object bounding box to object o_j, which is initialized as

$$(m,j) = \sum_{i=1}^{N_h}\sum_{l=1}^{L} p(H=h_i)\cdot\lambda^T_{i,j,l}\cdot b\left(\mathbf{x}^l_H, O^m\right) + \sum_{i=1}^{N_h}\sum_{k=1}^{N_a} p(H=h_i)\cdot p(A=a_k)\cdot\zeta_{i,j,k} + \gamma^T_j\cdot g(O^m) \quad (7)$$

Initializing the labels of all the windows to be background (i.e. no object in the bounding box), the forward search repeats the following steps

1) Select $(m^*, j^*) = \arg\max\{(m, j)\}$.
2) Label the m^*-th object detection window as o_{j^*} and remove it from the set of detection windows.
3) Update $(m, j) = (m, j) + \gamma^T_{j,j^*}\cdot b(O^m, O^{m^*}) + \gamma^T_{j^*,j}\cdot b(O^{m^*}, O^m)$.

until $(m^*, j^*) < 0$. After this step, all the bounding boxes are assigned to either an object or the background.

Updating the activity and atomic pose labels. Based on the current pose estimation and object detection results, we optimize $\Psi(A, O, H, I)$ by enumerating all possible combinations of A and H labels.

In this paper, we perform the above three steps for two iterations to obtain the inference results.

6 Experiments

In this section, we evaluate the performance of our method on two known datasets of HOI activities: a six-class sports dataset [14] and a 24-class people interacting with musical instrument (PPMI) dataset [15].

6.1 The Sports Dataset

The dataset. The sports dataset contains images of six sports activities. Instead of limiting to one human and one object interactions as in [50], here we consider all the objects that are involved in the activities. The objects that we consider are: cricket bat, ball, and stump in "cricket batting" and "cricket bowling"; croquet mallet, ball, and hoop in "croquet shot"; tennis racket and ball in "tennis forehand" and "tennis serving"; volleyball and net in "volleyball smash". These objects are either directly manipulated by humans, such as cricket bat and tennis racket, or highly related to the scene context of the activities, such as croquet hoop and volleyball net. There are 50 images in each activity class. We use the same setting as in [14]: 30 images for training and 20 for testing. In all the experiments of this section, all training phases are performed on this training set, including training the detectors for objects and body parts.

We train an upper-body detector on this dataset using the deformable part model [36] based on annotations on training images. The detector works almost perfectly on testing images because of the relatively clean image background. We normalize all the images based on the size of the detection boxes such that we do not need to search over different scales in human pose estimation.

Model visualization. On this dataset, there are six activities, ten objects, and twelve atomic poses. Conditioned on the image evidence (ϕ_{3-5} in Eq.1), our model learns the co-occurrence statistics (ϕ_1 in Eq.1) between these components as well as the spatial relationship (ϕ_2 in Eq.1) between objects and different body parts. Fig.7 visualizes the model parameters estimated from ϕ_1, i.e. the strength of the co-occurrence relationship between the set of activities, objects, and atomic poses. Each connection is obtained by marginalizing ζ in Eq.2 with respect to the other concept. For example, the strength of

TABLE 1
Object detection results on the sports dataset. We use average precision to measure the performance. The best results are marked by bold font in each experiment.

Method	Deformable model [36]	object context	person context	Our Method
cricket bat	17%	18%	**20%**	**20%**
cricket ball	24%	27%	25%	**32%**
cricket stump	77%	**78%**	**78%**	77%
croquet mallet	29%	32%	33%	**34%**
croquet ball	50%	52%	53%	**58%**
croquet hoop	15%	17%	18%	**22%**
tennis racket	33%	31%	32%	**37%**
tennis ball	42%	46%	45%	**49%**
volleyball	64%	65%	**67%**	**67%**
volleyball net	4%	6%	6%	**9%**
overall	36%	37%	38%	**41%**

TABLE 2
Human pose estimation results on the sports dataset. We use detection accuracy to measure the performance. "PS" stands for "pictorial structure". "Class-based PS" means training one pictorial structure using the images of each class, and apply the model to the testing images of the same class. The best results are marked by bold font.

Method	Yao & Fei -Fei [50]	PS [26]	Class -based PS	Our Method
head	58%	71%	70%	**76%**
torso	66%	69%	69%	**77%**
left/right upper arms	44%	44%	43%	**52%**
	40%	40%	42%	**45%**
left/right lower arms	27%	35%	37%	**39%**
	29%	36%	36%	**37%**
left/right upper legs	43%	58%	59%	**63%**
	39%	**63%**	62%	61%
left/right lower legs	44%	59%	**60%**	**60%**
	34%	71%	71%	**77%**
overall	42%	55%	56%	**59%**

the connection between pose h_i and object o_j is estimated by $\sum_{k=1}^{N_a} \exp(\zeta_{i,j,k})$.

Fig.7 shows that our method learns meaningful activity-pose-object interactions in HOI activities, such as the strong connection between 'tennis forehand' and the fourth atomic pose which is a reasonable gesture for the action, the object 'volleyball' and the last atomic pose, etc. Fig.7 also demonstrates the complex property of the interactions, as the interactions are not simple one-to-one mappings. Humans can manipulate the same object in more than one poses, while the same pose might correspond to many different objects. However, the human-object context does provide very useful information for understanding HOI activities. For example, it is more likely that the first atomic pose should connect to a cricket bat rather than a tennis racket. Fig.7 shows that our model successfully captures such information.

Object detection. One of our goals is to detect the presence and location of the objects involved in HOI activities. The experiment setting in this paper is different from that of [50] in two ways. Here we evaluate the performance of detecting each object in all the testing images, while in [50] only the images of the activity classes that could contain the object are considered (e.g. detecting volleyball in "volleyball smash" images). Furthermore in this paper we adopt a more standard experiment setting where the orientation of the objects are not considered.

We use the deformable part model [36] as the baseline for our object detection experiments. This is also the detection method used for initializing object detection scores in Eq.4 of our model. We further compare our method with two other control experiments with object context and person context respectively. For object context, we use the method in [10], where the spatial configurations between different objects serve as contextual information to improve the detection results. For person context, the upper-body detector provides the rough location of the human, which can be used to prune the object detection results lie in invalid geometric locations, e.g. cricket stump above the human. Detection performance is measured by average precision as in [9]. A detection bounding box is considered correct if the area of overlap between the detection box and the ground truth box exceeds 50% of the union of the two boxes. In our method, the confidence scores of the detection boxes are measured by Eq.7.

Table 1 shows the results of different methods. We observe that our detection method achieves the best performance. Compared with the other approaches without or with limited context, our method explores very detailed spatial relationship between different image parts, which helps to detect objects that are traditionally very difficult. For example, in the case of cricket ball and croquet ball, the deformable part model without context gives performance of 24% and 50%, while our method yields 32% and 58%. The reason might be that, on the one hand, the detailed human gesture helps to localize the balls which are almost impossible to detect by a simple sliding window approach, e.g. cricket ball in a human's hand when he is bowling the ball. On the other hand, our method distinguishes different sport activities (cricket batting or bowling versus croquet shot), hence is less likely to confuse the two types of balls.

Human pose estimation. Similarly to object detection, we show in this experiment that human pose estimation is significantly improved by object context. Here we compare our model with the state-of-the-art pictorial structure method [26]. We consider two approaches for training pictorial structures. One is to train a single model based on all the training images. The other is to train a pictorial structure model using training images of each activity class, and apply this model on the testing images of the same class.

Following the convention proposed in [58], a body

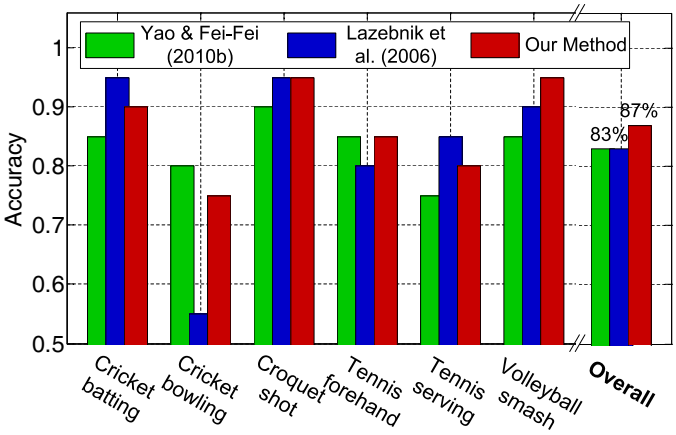


Fig. 8. Action classification results of different methods on the sports dataset. We use classification accuracy to evaluate the performance. "Overall" indicates the mean accuracy on all the classes.

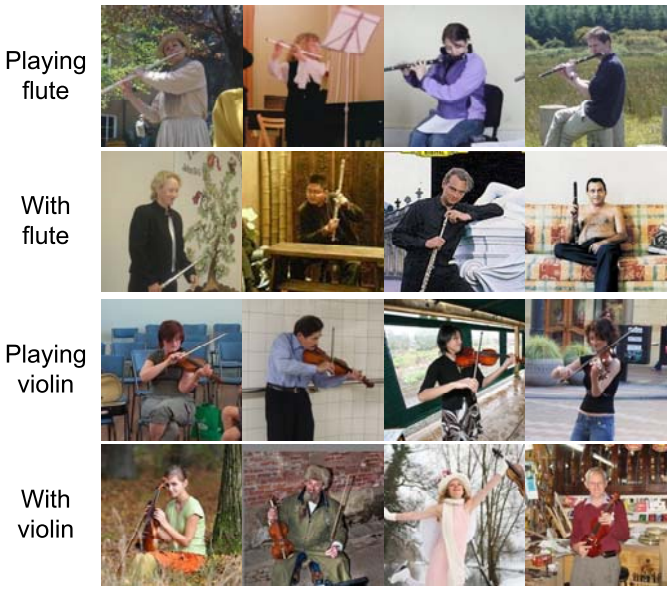


Fig. 9. An illustration of the difference between images of people playing musical instrument and people holding a musical instrument but not playing. In images of people just holding the instrument, both human poses and spatial relationships between humans and objects vary a lot.

part is considered correctly localized if the endpoints of its segment lie within 50% of the ground-truth segment length from their true positions. The missing body parts are not considered in performance evaluation. Experimental results are shown in Table 2. The percentage correctness tells us that pose estimation still remains a difficult problem, especially on the body parts that usually present large articulations and occlusions, such as lower arms. Our mutual context model outperforms the other approaches, even showing a 3% average improvement over a class-based pictorial structure model where uses the ground truth activity class labels. Furthermore, the performance of our mutual context model improves significantly comparing to [50], mainly because in this paper we normalize the testing images based on upper-body detection results and therefore searching over different scales of body parts is avoided.

Activity classification. Besides inferring the human pose and objects in the image, our model also gives a prediction of the class label of the HOI activity. Fig.8 compares the activity classification performance of our method and the results reported in [14], which also makes use of objects and human poses. We observe that our method outperforms [14] by 8%. Fig.8 shows that our method also outperforms the SPM approach, which we use to model the overall activity in Eq.6, demonstrating the effectiveness of human pose estimation and object detection for better activity classification. In Fig.11 we show examples of our object detection and pose estimation results in different activities.

6.2 The PPMI Dataset

The dataset. The PPMI dataset contains images of people interacting with twelve classes of musical instruments. Images of seven instruments bassoon, erhu, flute, French horn, guitar, saxophone, and violin were collected in [15]. Images of the other five instruments cello, clarinet, harp, recorder, and trumpet were later added. A very important property of this dataset is, for each musical instrument, there are images of people playing the musical instrument, as well as images of people holding the instrument but not playing it. Therefore there are 24 different human-object interactions in this dataset. This provides us the opportunity to analyze different interactions between humans and the same object, where the "playing" interaction usually corresponds to specific human poses and carries very special meanings, while the human gesture in the "not playing" interactions are relatively random (shown in Fig.9).

We use the "normalized images" of the dataset, where the humans interacting with the corresponding musical instruments are cropped and normalized to the same size (256×256 pixels) based on annotations of human faces. The full PPMI dataset contains 100 normalized training images and 100 normalized testing images for each interaction. In our experiment, we use a subset of the full dataset where each image contains only one person. Therefore for each class of interaction, we have $50\sim90$ images for training and the similar number of images for testing.

Model visualization. As on the sports dataset, we also obtain twelve atomic poses on the PPMI dataset. Our model learns the compatibility between these atomic poses and the set of activities and objects (ϕ_1 in Eq.1). For each musical instrument, the PPMI dataset contains two different interactions: playing and holding the instrument but not playing. Fig.10 shows the distribution of images of different interactions on the set of atomic poses. Intuitively speaking, humans usually play a musi-

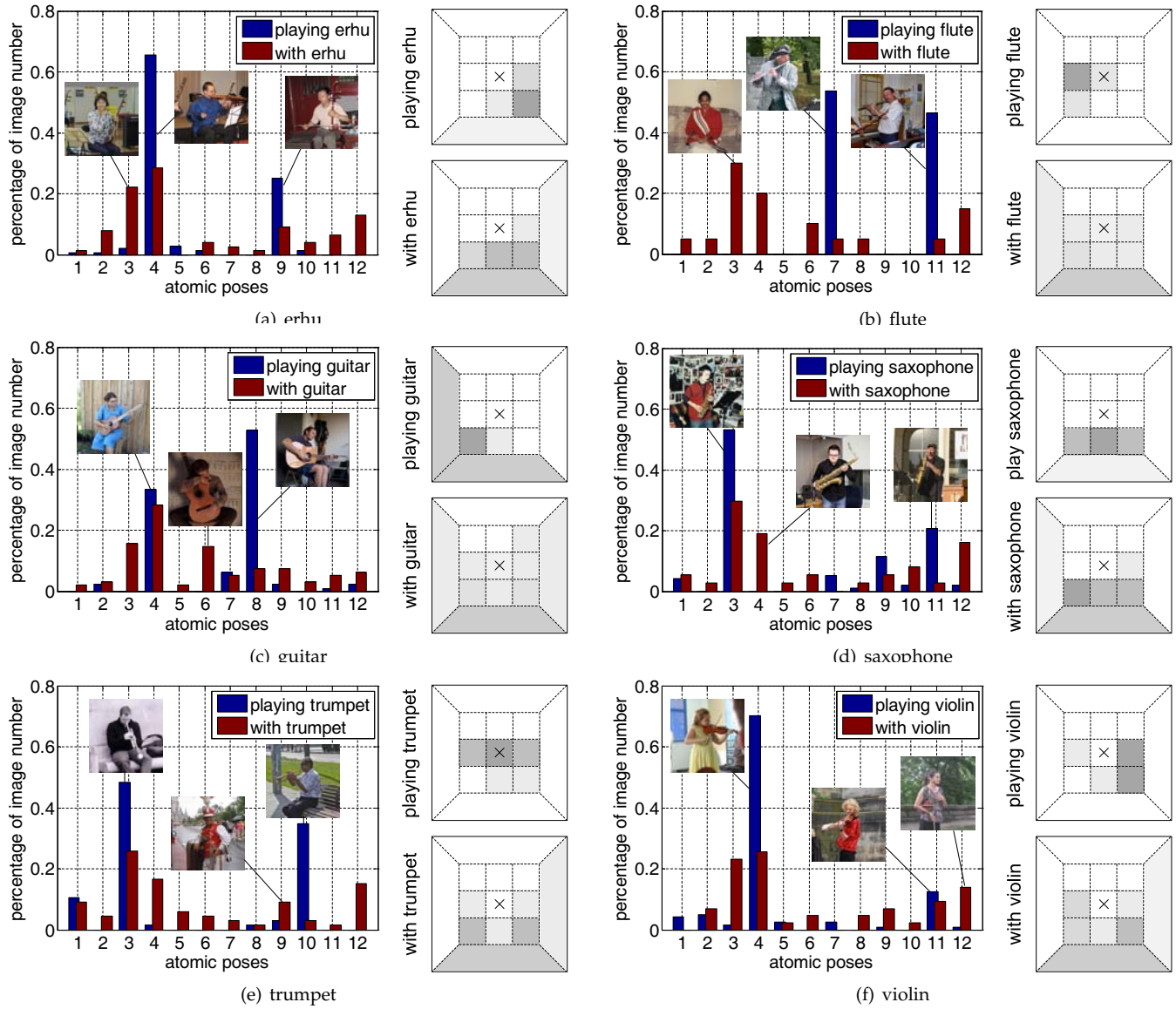


Fig. 10. Demonstration of our model on the PPMI dataset. For each instrument, the left bar figure illustrates the distribution of the number of images in each interaction on the twelve atomic poses. The digits in the horizontal axis indicate the atomic poses. Y-axis is the percentage of the number of images in the corresponding atomic pose. The right two images show the spatial relationship between the object and the human face. The location of the human face is denoted by "$\times$". Darker color indicates that the object is more likely to appear in the location. The upper image is for playing the instrument and the lower image is for holding the instrument but not playing.

cal instrument with some specific poses, while a human can hold an instrument in many possible poses when he is not playing it. This intuition is demonstrated in Fig.10. For each musical instrument, most images of people playing the instrument falls in a very small number of human poses, while the distributions of images of people not playing the instrument are more uniform. From Fig.10, we can also see that pose variation in the images of people playing the same musical instrument is mainly due to different shooting angles when the images are taken. Furthermore, we observe from Fig.10 that people might play different musical instruments with very similar poses. For example, images of playing erhu, guitar, and violin all have large distribution values on the fourth atomic pose.

Our approach also models the spatial relationship between objects and human body parts (ϕ_2 in Eq.1). We visualize in Fig.10 the spatial relationship between different objects and the human face. Our visualization is based on the division of image regions that we use in Eq.3. An image region is filled with a darker color if the likelihood for the corresponding object to be in that region is large. We compute the likelihood by marginalizing λ in Eq.3 with respect to all the atomic poses. For example, $\sum_{i=1}^{N_h} \exp(\lambda_{i,j,l}) \cdot p(h_i|a_k)$ measures the likelihood of the spatial relationships for object o_j

TABLE 3
Object detection results on the PPMI dataset. We evaluate the performance of different methods on images of people playing and not playing the instruments separately. As in Table 1, we use average precision to measure the performance. The best results are marked by bold font in each experiment.

Playing instrument:

Method	Deformable model [36]	person context	Our Method
bassoon	19%	21%	**24%**
cello	45%	47%	**52%**
clarinet	13%	15%	**19%**
erhu	32%	**35%**	**35%**
flute	23%	29%	**30%**
French horn	47%	48%	**49%**
guitar	48%	50%	**54%**
harp	33%	**34%**	**34%**
recorder	17%	18%	**24%**
saxophone	40%	44%	**45%**
trumpet	25%	27%	**32%**
violin	35%	39%	**41%**
overall	31%	34%	**37%**

Not playing instrument:

Method	Deformable model [36]	person context	Our Method
bassoon	**21%**	20%	**21%**
cello	45%	48%	**50%**
clarinet	16%	14%	**17%**
erhu	31%	35%	**36%**
flute	25%	29%	**31%**
French horn	**47%**	45%	46%
guitar	49%	**51%**	50%
harp	**34%**	32%	**34%**
recorder	16%	19%	**21%**
saxophone	39%	42%	**45%**
trumpet	26%	27%	**31%**
violin	38%	38%	**41%**
overall	32%	33%	**35%**

with respect to the l-th body part in action a_k, where $p(h_i|a_k)$ is the proportion of images of activity a_k in pose h_i. Similarly, we observe that the instruments fall in a small number of image regions in the interactions of "playing instrument", while the locations of the instruments are more random if the interaction is "not playing". Note that although we do not distinguish different activity classes in Eq.3, the difference between different interactions between humans and the same object can be captured by modeling different atomic poses in Eq.3.

Object detection. Here we test the performance of detecting the twelve musical instruments from the images. Since the playing and not playing are two interactions that carry different visual and functional meanings (Fig.10), we evaluate the object detection performance on images of the two interactions separately. Note that training of the instrument detectors and our model parameters are still done on all the images of 24 interactions. All the other experiment settings are the same as that on the sports dataset.

TABLE 4
Human pose estimation results on the PPMI dataset. We estimate the performance on images of people playing and not playing the musical instruments separately. As in Table 2, we use detection accuracy to measure the performance of localizing each body part in the images. The best results are marked by bold font.

Playing instrument:

Method	PS [26]	Class -based PS	Our Method
head	73%	**78%**	77%
torso	70%	**77%**	**77%**
left/right upper arms	40%	**43%**	42%
	37%	**42%**	41%
left/right lower arms	33%	38%	**40%**
	32%	**39%**	**39%**
overall	48%	**53%**	**53%**

Not playing instrument:

Method	PS [26]	Class -based PS	Our Method
head	74%	75%	**77%**
torso	**72%**	**72%**	71%
left/right upper arms	44%	45%	**46%**
	40%	**41%**	39%
left/right lower arms	36%	34%	**38%**
	36%	35%	**37%**
overall	50%	50%	**51%**

The average precision obtained from different approaches are listed in Table 3. There is only one object in each image that we use. So we do not compare our method with the approach that uses the spatial relationships between different objects [10]. Similar to the results on the sports dataset, our method largely outperforms the detector without context. Our method also performs better than the weak person context, because we have a more detailed and accurate modeling of the human pose as well as activity classes. Furthermore, we observe that our model does a better job in improving detection results on the images of people playing musical instruments, where the objects are manipulated by the humans in some specific ways (shown in Fig.9) and therefore more structural information can be discovered. Specifically, our method obtains 7% performance gain in detecting cello, flute, and recorder in the interactions of people playing musical instruments.

Human pose estimation. Here we evaluate the performance of human pose estimation on images of people playing and not playing musical instruments separately. Experimental results are shown in Table 4. Similarly to the object detection results, we observe that on the images of people playing musical instruments, our method performs much better than a single pictorial structure model. But our method is only slightly better than the pictorial structure model on the images where people

TABLE 5
Activity classification results on the PPMI dataset. The best performance is marked by bold font.

Method	SPM [56]	Grouplet [15]	Our Method
play bassoon	37%	30%	**42%**
play cello	41%	41%	**50%**
play clarinet	39%	43%	**49%**
play erhu	48%	43%	**54%**
play flute	41%	47%	**53%**
play French horn	44%	38%	**52%**
play guitar	40%	50%	**52%**
play harp	44%	36%	**45%**
play recorder	45%	49%	**56%**
play saxophone	42%	**49%**	47%
play trumpet	39%	**53%**	47%
play violin	43%	50%	**51%**
with bassoon	38%	41%	**47%**
with cello	42%	32%	**54%**
with clarinet	39%	39%	**48%**
with erhu	35%	35%	**41%**
with flute	48%	**49%**	45%
with French horn	36%	**53%**	43%
with guitar	34%	41%	**42%**
with harp	38%	33%	**47%**
with recorder	42%	**52%**	**52%**
with saxophone	36%	36%	**50%**
with trumpet	39%	43%	**48%**
with violin	35%	30%	**45%**
overall	40%	42%	**48%**

just hold the musical instruments but not playing. We also observe that training a pictorial structure model for each class performs on par with our mutual context model, which shows the consistency of human poses when humans are playing musical instruments (shown in Fig.9). However, the ground truth labels of activity classes are used when testing the performance of class-based pictorial structure, while our method is fully automatic in the testing phase.

Activity classification. Table 5 shows the activity classification result on the PPMI dataset. By jointly modeling the objects and human poses, our method outperforms the spatial pyramid matching [56] and grouplet [15] approaches by a large margin.

7 Conclusion

In this work, we treat object and human pose as the context of each other in different HOI activity classes. We develop a conditional random field model that learns co-occurrence context and spatial context between objects and human poses. Experimental results show that our model significantly outperforms other state-of-the-art methods in both problems.

One major contribution of this work is to demonstrate the importance of context in visual recognition. Specifically, we study a new problem (recognizing human-object interaction activities) where context between objects and human poses can significantly improve recognition performance. Nevertheless, there are many other scenarios where context plays critical roles, e.g. detecting the keyboard and mouse near a computer monitor. It would be worthwhile to design computer vision techniques that make use of context in such situations.

One limitation of our work is we need to annotate the human body parts and objects in each training image. One direction of our future work is to study weakly supervised or unsupervised approaches to understand human-object interaction activities.

Acknowledgments

L.F-F. is partially supported by an NSF CAREER grant (IIS-0845230), an ONR MURI grant, the DARPA VIRAT program, the DARPA Mind's Eye program, Intel ISTC-PC, a Google research award, and a Microsoft Research Fellowship. B.Y. is partially supported by the SAP Stanford Graduate Fellowship. We would like to thank Nityananda J for help annotating human body parts in the PPMI dataset. We also would like to thank all the external reviewers for helpful suggestions to this work.

References

[1] I. Biederman, R. Mezzanotte, and J. Rabinowitz, "Scene perception: Detecting and judging objects undergoing relational violations," *Cognitive Psychology*, vol. 14, pp. 143–177, 1982.
[2] A. Oliva and A. Torralba, "The role of context in object recognition," *Trends in Cognitive Sciences*, vol. 11, no. 12, pp. 520–527, 2007.
[3] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie, "Objects in context," in *International Conference on Computer Vision*, 2007.
[4] G. Heitz and D. Koller, "Learning spatial context: Using stuff to find things," in *European Conference on Computer Vision*, 2008.
[5] S. Divvala, D. Hoiem, J. Hays, A. Efros, and M. Hebert, "An empirical study of context in object detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.
[6] K. Murphy, A. Torralba, and W. Freeman, "Using the forest to see the trees: a graphical model relating features, objects, and scenes," in *Advances in Neural Information Processing Systems*, 2003.
[7] M. Marszalek, I. Laptev, and C. Schmid, "Actions in context," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.
[8] J. Shotton, J. Winn, C. Rother, and A. Criminisi, "TextonBoost: joint appearance, shape and context modeling for multi-class object recognition and segmentation," in *European Conference on Computer Vision*, 2006.
[9] M. Everingham, L. V. Gool, C. Williams, J. Winn, and A. Zisserman, "The PASCAL VOC2008 Results."
[10] C. Desai, D. Ramanan, and C. Fowlkes, "Discriminative models for multi-class object layout," in *International Conference on Computer Vision*, 2009.
[11] H. Harzallah, F. Jurie, and C. Schmid, "Combining efficient object localization and image classification," in *International Conference on Computer Vision*, 2009.
[12] B. Leibe, A. Leonardis, and B. Schiele, "Combined object categorization and segmentation with an implicit shape model," in *ECCV Workshop on Statistical Learning in Computer Vision*, 2004.
[13] J. Henderson, "Human gaze control during real-world scene perception," *Trends in Cognitive Sciences*, vol. 7, no. 11, pp. 498–504, 2003.
[14] A. Gupta, A. Kembhavi, and L. Davis, "Observing human-object interactions: Using spatial and functional compatibility for recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 10, pp. 1775–1789, 2009.
[15] B. Yao and L. Fei-Fei, "Grouplet: A structured image representation for recognizing human and object interactions," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

Fig. 11. Example testing results of object detection and human pose estimation in different HOI activities. The color codes for the objects and different body parts are: objects - red, head and torso - magenta, arms - green, legs - cyan.

[16] B. Yao, A. Khosla, and L. Fei-Fei, "Classifying actions and measuring action similarity by modeling the mutual context of objects and human poses," in *International Conference on Machine Learning*, 2011.

[17] D. Bub and M. Masson, "Gestural knowledge evoked by objects as part of conceptual representations," *Aphasiology*, vol. 20, pp. 1112–1124, 2006.

[18] H. Helbig, M. Graf, and M. Kiefer, "The role of action representation in visual object," *Experimental Brain Research*, vol. 174, pp. 221–228, 2006.

[19] P. Bach, G. Knoblich, T. Gunter, A. Friederici, and W. Prinz, "Action comprehension: Deriving spatial and functional relations," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 31, no. 3, pp. 465–479, 2005.

[20] A. Efros, A. Berg, G. Mori, and J. Malik, "Recognizing action at a distance," in *International Conference on Computer Vision*, 2003.

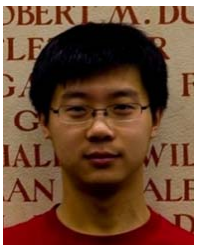
[21] I. Laptev, "On space-time interest points," *International Journal of Computer Vision*, vol. 64, no. 2/3, pp. 107–123, 2005.

[22] J. Liu, J. Luo, and M. Shah, "Recognizing realistic actions from videos "in the wild"," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[23] J. Niebles, C. Chen, and L. Fei-Fei, "Modeling temporal structure of decomposable motion segments for activity classification," in *European Conference on Computer Vision*, 2010.

[24] P. Felzenszwalb and D. Huttenlocher, "Pictorial structures for object recognition," *International Journal of Computer Vision*, vol. 61,

no. 1, pp. 55–79, 2005.
[25] D. Ramanan, "Learning to parse images of articulated objects," in *Advances in Neural Information Processing Systems*, 2006.
[26] M. Andriluka, S. Roth, and B. Schiele, "Pictorial structures revisited: People detection and articulated pose estimation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.
[27] M. Eichner and V. Ferrari, "Better appearance models for pictorial structures," in *British Machine Vision Conference*, 2009.
[28] B. Sapp, A. Toshev, and B. Taskar, "Cascade models for articulated pose estimation," in *European Conference on Computer Vision*, 2010.
[29] Y. Yang and D. Ramanan, "Articulated pose estimation with flexible mixture-of-parts," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.
[30] X. Ren, A. Berg, and J. Malik, "Recovering human body configurations using pairwise constraints between parts," in *International Conference on Computer Vision*, 2005.
[31] Y. Wang and G. Mori, "Multiple tree models for occlusion and spatial constraints in human pose estimation," in *European Conference on Computer Vision*, 2008.
[32] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
[33] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.
[34] P. Viola and M. Jones, "Robust real-time object detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2001.
[35] C. Lampert, M. Blaschko, and T. Hofmann, "Beyond sliding windows: object localization by efficient subwindow search," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.
[36] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.
[37] D. Hoiem, A. Efros, and M. Hebert, "Putting objects in perspective," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2006.
[38] A. Gupta, T. Chen, F. Chen, D. Kimber, and L. Davis, "Context and observation driven latent variable model for human pose estimation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.
[39] H. Kjellstrom, D. Kragic, and M. Black, "Tracking people interacting with objects," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.
[40] B. Rosenhahn, C. Schmaltz, T. Brox, J. Weickert, and H.-P. Seidel, "Staying well grounded in markerless motion capture," in *Symposium of the German Association for Pattern Recognition*, 2008.
[41] M. Brubaker, L. Sigal, and D. Fleet, "Estimating contact dynamics," in *International Conference on Computer Vision*, 2009.
[42] C. Desai, D. Ramanan, and C. Fowlkes, "Discriminative models for static human-object interactions," in *CVPR Workshop on Statistical Models in Computer Vision*, 2010.
[43] W. Yang, Y. Wang, and G. Mori, "Recognizing human actions from still images with latent poses," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.
[44] V. Delaitre, I. Laptev, and J. Sivic, "Recognizing human actions in still images: A study of bag-of-features and part-based representations," in *British Machine Vision Conference*, 2010.
[45] S. Maji, L. Bourdev, and J. Malik, "Action recognition from a distributed representation of pose and appearance," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.
[46] A. Prest, C. Schmid, and V. Ferrari, "Weakly supervised learning of interaction between humans and objects," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. 99, 2011.
[47] L. Jie, B. Caputo, and V. Ferrari, "Who's doing what: Joint modeling of names and verbs for simultaneous face and pose annotation," in *Advances in Neural Information Processing Systems*, 2009.
[48] V. Singh, F. Khan, and R. Nevatia, "Multiple pose context trees for estimating human pose in object context," in *CVPR Workshop on Structural Models in Computer Vision*, 2010.
[49] M. Sadeghi and A. Farhadi, "Recognition using visual phrases," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.
[50] B. Yao and L. Fei-Fei, "Modeling mutual context of object and human pose in human-object interaction activities," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.
[51] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *International Conference on Machine Learning*, 2001.
[52] J. Liebelt, C. Schmid, and K. Schertler, "Viewpoint-independent object class detection using 3D feature maps," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.
[53] Y. Wang, H. Jiang, M. Drew, Z.-N. Li, and G. Mori, "Unsupervised discovery of action classes," in *CVPR*, 2006.
[54] L. Bourdev and J. Malik, "Poselets: Body part detectors trained using 3D human pose annotations," in *International Conference on Computer Vision*, 2009.
[55] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2005.
[56] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2006.
[57] D. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.
[58] V. Ferrari, M. Marín-Jiménez, and A. Zisserman, "Progressive search space reduction for human pose estimation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.



Bangpeng Yao is a PhD student in the Computer Science Department at Stanford University. He received his B.E. degree in Automation and a M.E. degree in Computer Science from Tsinghua University in 2006 and 2008 respectively, both with honors. He was a PhD student in the Computer Science Department of Princeton University for one year, and then moved to Stanford in 2009. His research interests are in computer vision and applied machine learning. He is a recipient of the Stanford Graduate Fellowship for 2010-2012, and the Microsoft Research PhD Fellowship for 2012-2013. In 2010, he received the CVPR Best Paper Honorable Mention award.

Li Fei-Fei is an assistant professor in the Computer Science Department at Stanford University and director of the Stanford Vision Lab. She also holds courtesy appointments in the Neuroscience Program and the Psychology Department at Stanford. Her main research interest is in vision, particularly high-level visual recognition. In computer vision, her interests span from object and natural scene understanding to activity and event recognition in both videos and still images. In human vision, she and her students have studied the interaction of attention and natural scene and object recognition, and decoding the human brain fMRI activities involved in natural scene categorization by using pattern recognition algorithms.

She received her A.B. degree in physics from Princeton University, and subsequently her Ph.D. degree in electrical engineering from the California Institute of Technology. From 2005 to August 2009, she was an assistant professor in the Electrical and Computer Engineering Department at University of Illinois Urbana-Champaign and Computer Science Department at Princeton University, respectively. She joined Stanford in 2009. She has published over 60 peer-reviewed papers in computer vision, cognitive neuroscience and machine learning at top journals and conferences. She is a recipient of a Microsoft Research New Faculty award, an NSF CAREER award, two Google Research Awards, and a CVPR Best Paper Honorable Mention award.