

Modeling Mutual Context of Object and Human Pose in Human-Object Interaction Activities

Bangpeng Yao Li Fei-Fei

Computer Science Department, Stanford University, USA

{bangpeng, feifeili}@cs.stanford.edu

Abstract

Detecting objects in cluttered scenes and estimating articulated human body parts 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 object tends 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 new random field model to encode the mutual context of objects and human poses in human-object interaction activities. We then cast the model learning task as a structure learning problem, of which the structural connectivity between the object, the overall human pose, and different body parts are estimated through a structure search approach, and the parameters of the model are estimated by a new max-margin algorithm. On a sports data set of six classes of human-object interactions [12], we show that our mutual context model significantly outperforms state-of-the-art in detecting very difficult objects and human poses.

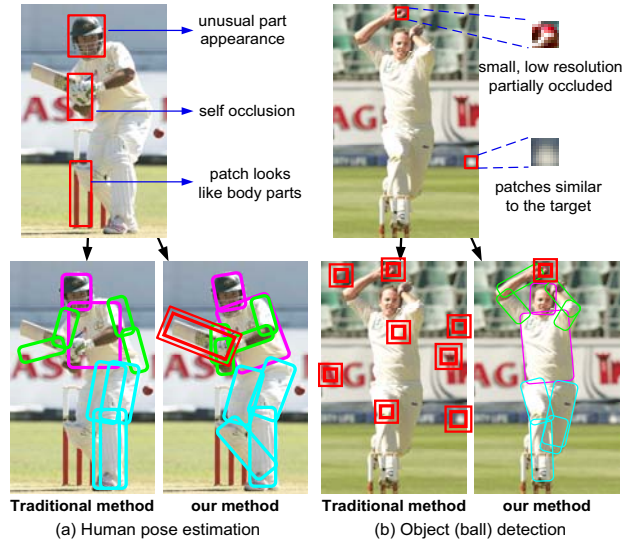


Figure 1. Objects and human poses can serve as mutual context to facilitate the recognition of each other. In (a), the human pose is better estimated by seeing the cricket bat, from which we can have a strong prior of the pose of the human. In (b), the cricket ball is detected by understanding the human pose of throwing the ball.

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 [3, 24]. In computer vision, context has been used in problems such as object detection and recognition [25, 14, 8], scene recognition [23], action classification [22], and segmentation [28]. 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 dataset [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% [7, 13].

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 [20] can nevertheless detect cars with high accuracy whether they 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) [15].

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 [12, 32]. As shown in Fig. 1, 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 to detect the small ball in the player’s hand, which is nearly

invisible even to the human eye (Fig. 1(b)).

However, the two difficult tasks can benefit greatly from serving as context for each other, as shown in Fig. 1. The goal of this paper is to model the *mutual context* of objects and human poses in HOI activities so that each can facilitate the recognition of the other. Given a set of training images, our model automatically discovers the relevant poses for each type of HOI activity, and furthermore the connectivity and spatial relationships between the objects and body parts. We formulate this task as a structure learning problem, of which the connectivity is learned by a structure search approach, and the model parameters are discriminatively estimated by a novel max-margin approach. By modeling the mutual co-occurrence and spatial relations of objects and human poses, we show that our algorithm significantly improves the performance of both object detection and pose estimation on a dataset of sports images [12].

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.

2. Related work

The two central tasks, human pose estimation and object detection, have been studied in computer vision for many years. Most of the pose estimation work uses a tree structure of the human body [10, 26, 1] which allows fast inference. In order to capture more complex body articulations, some non-tree models have also been proposed [27, 31]. Although those methods have been demonstrated to work well on the images with clean backgrounds, human pose estimation in cluttered scenes remains a challenging problem. Furthermore, to our knowledge, no existing method has explored context information for human pose estimation.

Sliding window is one of the most successful strategies for object detection. Some techniques have been proposed to avoid exhaustively searching the image [30, 19], which makes the algorithm more efficient. While the most popular detectors are still based on sliding windows, more recent work has tried to integrate context to obtain better performance [25, 14, 8]. 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 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 [23, 17, 28, 25, 14, 8, 22]. In most of these works, 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. In this paper, we try to bridge the gap between two seemingly unrelated problems - object detection and human pose



(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 way that the human interacts with the object (the right two images).

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

estimation, in which the *mutual contexts* play key roles for understanding their interactions. The problem of classifying HOI activities has been studied in [12] and [32], but no detailed understanding of the human pose (e.g. parsing the body parts) is offered in these works. To our knowledge, our work is the first one that explicitly models the mutual contexts of human poses and objects and allows them to facilitate the recognition of each other.

3. Modeling mutual context of object and pose

Given an HOI activity, our goal is to estimate the human pose and to detect the object that the human interacts with. Fig. 2 illustrates that both tasks are challenging. 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 and many body parts are self-occluded. Furthermore, even in the same activity, the configurations of body parts might differ in different images due to different shooting angles or human poses.

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. For simplicity, we assume that only one object is involved in each activity.

3.1. The model

A graphical illustration of our model is shown in Fig. 3(a). Our model can be thought of as a hierarchical random field, where the overall activity class A , object O , and human pose H all contribute to the recognition and detection of each other. The human pose is further decomposed into some body parts, denoted by $\{P_n\}_{n=1}^N$. For each body part P_n and the object O , f_{P_n} and f_O denote the visual features that describe the corresponding image regions respectively. Note that because of the difference between

the human poses in each HOI activity (Fig.2(b)), we allow each activity class (A) to have more than one types of human pose (H), which are latent (unobserved) variables to be learned in training.

Our model encodes the mutual connections between the object, the human pose and the body parts. Intuitively speaking, this allows the model to capture important connections between, say, the tennis racket and the right arm that is serving the tennis ball (Fig.3(b)). We observe, however, that the left leg in tennis serving is often less relevant to the detection of the ball. The model should therefore have the flexibility in deciding what parts of the body should be connected to the object O and the overall pose H . Dashed lines in Fig.3(a) indicate that these connections will be decided through structure learning. Depending on A , O and H , these connections might differ in different situations. Putting everything together, the overall model can be computed as $\Psi = \sum_e w_e \psi_e$, where e is an edge of the model, ψ_e and w_e are its potential function and weight respectively. We now enumerate the potentials of this model:

- $\psi_e(A, O)$, $\psi_e(A, H)$, and $\psi_e(O, H)$ model the agreement between the class labels of A , O , and H , each estimated by counting the co-occurrence frequencies of the pair of variables on training images.

- $\psi_e(O, P_n)$ models the spatial relationship between the object O and the body part P_n , which is computed by

$$\text{bin}(\mathbf{l}_O - \mathbf{l}_{P_n}) \cdot \text{bin}(\theta_O - \theta_{P_n}) \cdot \mathcal{N}(s_O / s_{P_n}) \quad (1)$$

where $(\mathbf{l}, \theta, s)$ is the position, orientation, and scale of an image part. $\text{bin}(\cdot)$ is a binning function as in [26] and $\mathcal{N}(\cdot)$ is a Gaussian distribution.

- $\psi_e(P_m, P_n)$ models the spatial relationship between different body parts, computed similarly to Eq.1.
- $\psi_e(H, P_n)$ models the compatibility between the pose class H and a body part P_n . It is computed by considering the spatial layout of P_n given a reference point in the image, in this case the center of the human face (P_1).

$$\psi_e(H, P_n) = \text{bin}(\mathbf{l}_{P_n} - \mathbf{l}_{P_1}) \cdot \text{bin}(\theta_{P_n}) \cdot \mathcal{N}(s_{P_n}) \quad (2)$$

- $\psi_e(O, f_O)$ and $\psi_e(P_n, f_{P_n})$ model the dependence of the object and a body part with their corresponding image evidence. We use the shape context [2] feature for image representation, and train a detector [30] for each body part and each object in each activity. Detection outputs are normalized as in [1].

In our algorithm, all the above potential functions are dependent on O and H except those between A , O , and H (the first bullet). We omit writing this point every time for space consideration. For example, for different human pose H , $\psi_e(O, P_n)$ is estimated with different parameters, which represents a specific spatial configuration between P_n and the object O , conditioned on the particular human pose H .

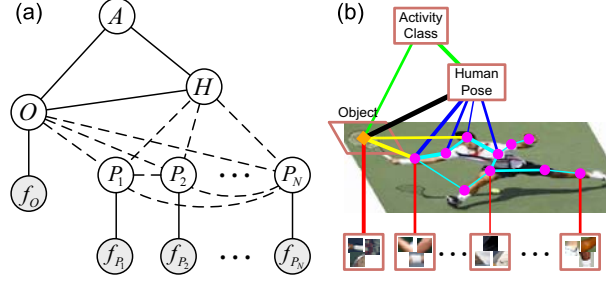


Figure 3. (a) A graphical illustration of our model. The edges represented by dashed lines indicate that their connectivity will be obtained by structure learning. A denotes an HOI activity class, H the human pose class, P a body part, and O the object. f_O and f_P 's are image appearance information of O and P respectively. (b) Illustration of our model on an image of a human playing tennis. Different types of potentials are denoted by lines with different colors. Line widths represent the importance of the potentials for the human-object interaction of playing tennis.

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. 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. 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. That is to say, co-occurrence information can be beneficial for coherently modeling the object, the human pose, and the activity class.

Multiple types of human poses for each activity. Our model allows each activity (A) to consist of more than one human pose (H). Treating H as a hidden variable, our model automatically discovers the possible poses from training images. This gives us more flexibility to deal with the situations where the human poses in the same activity are inconsistent, as shown in Fig.2(b). We show in Fig.4 the pose variability for each HOI activity.

Spatial context between object and body parts. Different poses imply that the object is handled by the human in different manners, which are modeled by $\{\psi_e(O, P_n)\}_{n=1}^N$. Furthermore, not all these relationships are critical for understanding an HOI activity. Therefore for each combination of O and H , our algorithm automatically discovers the connectivity between O and each P_n , as well as the connectivity among H and $\{P_n\}_{n=1}^N$.

Relations with the other models. Our model has drawn inspirations from a number of previous works, such as mod-

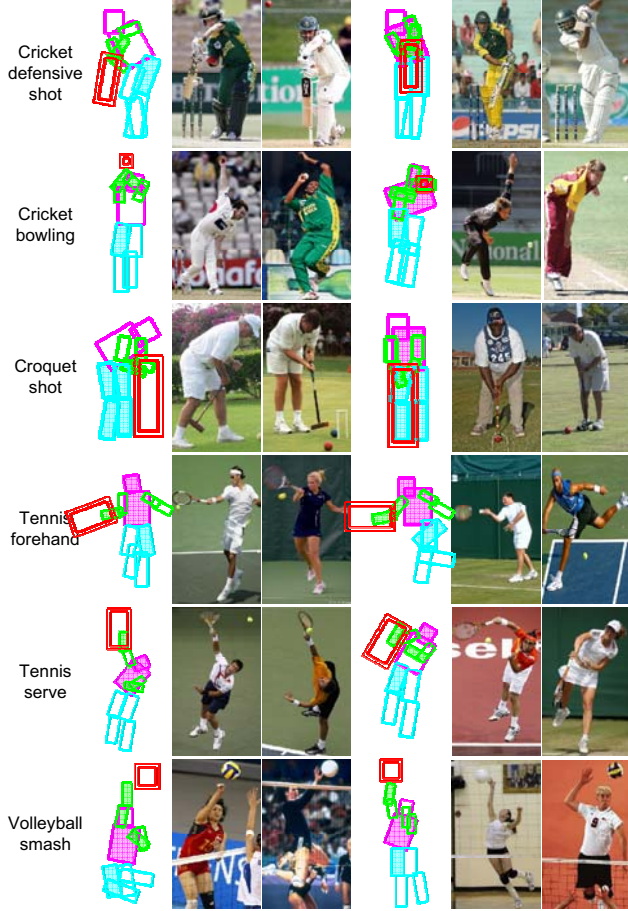


Figure 4. Visualization of the learned HOI models. Each row shows two models (due to two types of poses) and their corresponding image examples for one activity. The illustrative figure for each model represents the average spatial layout of the object and body parts of all the images that are assigned to the model. Shaded half-transparent body parts indicate that the structure learning algorithm has assigned strong connectivity to these parts with respect to the object. The different color codes are: object = double red box, head and torso = magenta, arms = green, legs = cyan.

eling spatial layout of different image parts [10, 26, 1], using agreement of different image components [25], using multiple models to describe the same concept (human pose in our problem) [21], non-tree models for better pose estimation [31, 4], and discriminative training [1]. Our model integrates all the properties in one coherent framework to perform two seemingly different tasks, human pose estimation and object detection, to the benefit of each other.

4. Model learning

Given the training images of HOI activities with labeled objects and body parts, the learning step needs to achieve two goals: *structure learning* to discover the hidden human

Hill-climbing structure learning for each activity class.

foreach *Iteration* **do**

- Model parameter estimation by max-margin learning;
- Choose the model with the largest number of mis-classified images;
- Cluster the images in the selected model into two sub-classes;
- Structure learning for the two new sub-classes;

end

Algorithm 1: The learning framework. Each sub-class corresponds to a type of human pose in an HOI activity. Initially there are one sub-class for each activity.

poses and the connectivity among the object, human pose, and body parts; and *parameter estimation* for the potential weights to maximize the discrimination between different activities. The output of our learning method is a set of models, each representing one connectivity pattern and potential weights for one type of human pose in one activity class. Algorithm 1 is a sketch of the overall framework. We discover new human poses by clustering the samples in the model that has the weakest discriminative ability in each iteration, which results to some sub-classes. Structure learning is applied to each sub-class respectively. The learning process terminates when the number of mis-classified samples in each sub-class is small (less than three in this paper).

4.1. Hill-climbing structure learning

Our algorithm performs structure learning for each sub-class, i.e. each pose in each activity, respectively. Given the images of a sub-class which are obtained from clustering (see Algorithm 1), our objective is to learn a connectivity pattern between the object, the human pose, and the body parts (the dashed lines in Fig.3(a)). Here we omit the edges between A , O , and H because they do not affect the structure learning results.

We use a hill-climbing approach with tabu list [18] to search the structure space. The hill-climbing approach adds or removes edges one at a time until a maximum is reached. During the search procedure, we keep a tabu list of history operations to guarantee that those operations will not be reversed. As in [14], we include a Gaussian prior over the number of edges to avoid overfitting. Furthermore, in order to reduce the impact of local maxima, we randomly initialize the structure for three times and apply the hill-climbing approach to each one, from which the best result is selected. Please find more details of our structure learning approach in Sec.A.

4.2. Max-margin parameter estimation

Given the model outputs by the structure learning step, the parameter estimation step aims to obtain a set of po-

tential weights that maximize the discrimination between different classes of activities (A in Fig.3(a)). But unlike the traditional random field parameter estimation setting [29], in our model each class can contain more than one pose (H), which can be thought of as multiple sub-classes. Our learning algorithm needs to, therefore, estimate parameters for each pose (i.e. sub-class) while optimizing for maximum discrimination among the global activity classes.

We propose a novel max-margin learning approach to tackle this problem. Let $(\mathbf{x}_i, c_i, y(c_i))$ be a training sample, where $\mathbf{x}_i$ is a data point, c_i is the sub-class label of $\mathbf{x}_i$, and $y(c_i)$ maps c_i to a class label. We want to find a function $\mathcal{F}$ that assigns an instance $\mathbf{x}_i$ to a sub-class. We say that $\mathbf{x}_i$ is correctly classified if and only if $y(\mathcal{F}(\mathbf{x}_i)) = y(c_i)$. Our classifier is then formulated as $\mathcal{F}(\mathbf{x}_i) = \arg \max_r \{\mathbf{w}_r \cdot \mathbf{x}_i\}$, where $\mathbf{w}_r$ is a weight vector for the r -th sub-class. Inspired by the traditional max-margin learning problems [5], we introduce a slack variable ξ_i for each sample $\mathbf{x}_i$, and optimize the following objective function:

$$\min_{\mathbf{w}, \xi} \frac{1}{2} \sum_r \|\mathbf{w}_r\|_2^2 + \beta \sum_i \xi_i \quad (3)$$

subject to: $\forall i, \xi_i \geq 0$

$\forall i, r$ where $y(r) \neq y(c_i)$, $\mathbf{w}_{c_i} \cdot \mathbf{x}_i - \mathbf{w}_r \cdot \mathbf{x}_i \geq 1 - \xi_i$

where $\|\mathbf{w}_r\|_2$ is the L2 norm of $\mathbf{w}_r$, β is a normalization constant. Again, note that the weights are defined with respect to sub-classes while the classification results are measured with respect to classes. We optimize Eq.3 by using the multiplier method [16]. Mapping the above symbols to our model, $\mathbf{x}_i$ are the potential function values computed on an image. Potential values for the disconnected edges are set to 0. In order to obtain better discrimination among different classes, we compute the potential values of an image on the models of all the sub-classes, and concatenate these values to form the feature vector. Sub-class variable c_i indicates human pose H , and $y(c_i)$ is the class label A . Please refer to Sec.B for more detail about the method.

4.3. Analysis of our learning algorithm

Fig.4 illustrates the two models (correspond to two types of human poses) learned by our algorithm for each HOI class. We can see the big difference of human poses in some activities (e.g. croquet-shot and tennis-serve), and such wide intra-class variability can be effectively captured by our algorithm. In these cases, using only one human pose for each HOI class is not enough to characterize well all the images in this class. Furthermore, we observe that by using structure learning, our model can learn meaningful connectivity between the object and the body parts, e.g. croquet mallet and legs, right forehand and tennis racket.

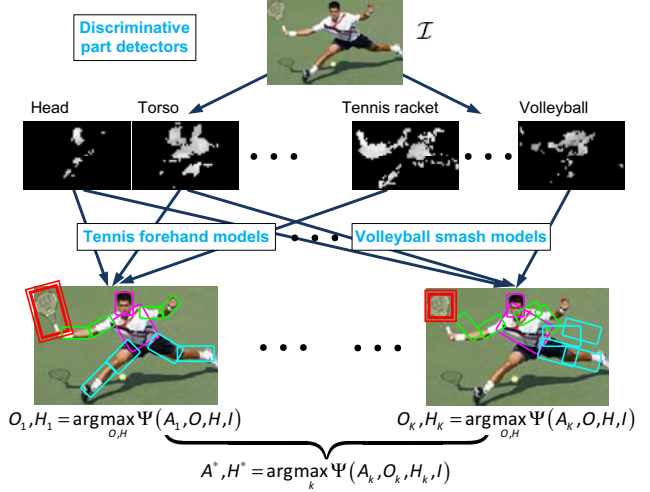


Figure 5. The framework of our inference method. Given an input image $\mathcal{I}$, the inference results are: (1) object detection results O_k (e.g. O_1 is the tennis racket detection result); (2) human pose estimation result H^* ; (3) activity classification result A^* .

5. Model inference, object detection, and human pose estimation

Given a new testing image $\mathcal{I}$, our objective is to estimate the pose of the human in the image, and to detect the object that is interacting with the human. An illustration of the inference procedure is shown in Fig.5. In order to detect the tennis racket in this image, we maximize the likelihood of this image given the models that are learned for tennis-forehand. This is achieved by finding a best configuration of human body parts and the object (tennis racket) in the image, which is denoted as $\max_{O, H} \Psi(A_k, O, H, \mathcal{I})$ in Fig.5. In order to estimate the human pose, we compute $\max_{O, H} \Psi(A_k, O, H, \mathcal{I})$ for each activity class and find the class A^* that corresponds to the maximum likelihood score. This score can be used to measure the confidence of activity classification as well as human pose estimation.

For each model, the above inference procedure involves a step to find the best spatial configuration of the object and different body parts for an image. We solve this problem by using the compositional inference method [4]. The inference algorithm contains a bottom-up stage to make proposals for the image parts. Each bottom-up step is followed by a top-down stage to validate the proposals. More details of the compositional inference method can be found in Sec.C of this paper.

6. Experiments

6.1. The sports dataset

We evaluate our approach on a known HOI dataset of six activity classes [12]: cricket-defensive shot (player

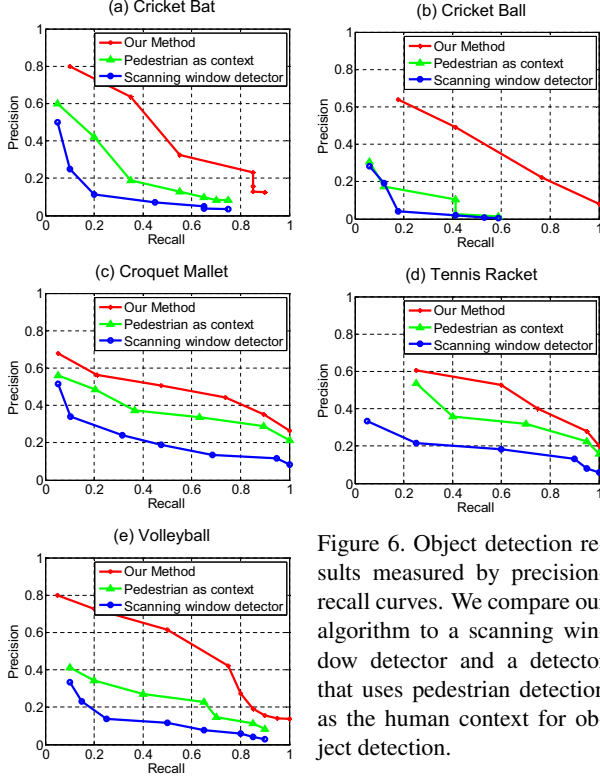


Figure 6. Object detection results measured by precision-recall curves. We compare our algorithm to a scanning window detector and a detector that uses pedestrian detection as the human context for object detection.

and cricket bat), cricket-bowling (player and cricket ball), croquet-shot (player and croquet mallet), tennis-forehand (player and tennis racket), tennis-serve (player and tennis racket), and volleyball-smash (player and volleyball). There are 50 images in each activity class. We use the same setting as that in [12]: 30 images for training and 20 for testing. In [12] only activity classification results were reported. In this work we also evaluate our method on the tasks of object detection and human pose estimation.

6.2. Better object detection by pose context

In this experiment, our goal is to detect the presence and location of the object given an HOI activity. To evaluate the effectiveness of our model, we compare our results with two control experiments: a scanning window detector as a baseline measure of object detection without any context, and a second experiment in which the approximate location of the person is provided by a pedestrian detector [6], hence providing a co-occurrence context and a very weak location context. Results of these three experiments, measured by precision-recall curves, are shown in Fig.6. The curves of our algorithm are obtained by considering the scores $\Psi(A, O, H, \mathcal{I})$ of all the results that are proposed by the compositional inference method. To ensure fair comparison, all experiments use the same input features and object detectors described in Sec.3, and non-max suppression is applied equally to all methods.

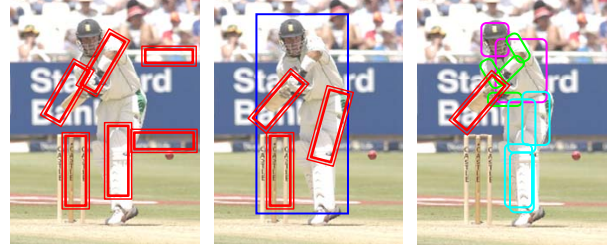


Figure 7. Object (cricket bat) detection results (red double-line bounding boxes) obtained by: a sliding window detector (left), the same detector using pedestrian detection as context (middle), and our method (right). Pedestrian detection is shown in a blue bounding box. The human pose estimation results are shown in colored rectangles in the right image.

The results in Fig.6 show that our detection method achieves the best performance. By using human pose as context, more detailed spatial relationship between different image parts can be discovered, which greatly helps to detect objects that are traditionally very difficult. For example, in the case of the cricket ball (Fig.6(b)), a sliding window method yields an average precision of 17%, whereas our model with pose-context measure is 46%. In almost all the cases of the five objects, the average precision score of our method is more than three times as the sliding window method. Fig.7 shows an example of using the three methods for object detection. Fig.9 shows more object detection results on a variety of testing images.

6.3. Better pose estimation by object context

Similarly to object detection, we show in this experiment that human pose estimation is significantly improved by object context. Here we compare our full model with four different control experiments.

- An *iterative parsing* method by Ramanan et al [26];
- A state-of-the-art *pictorial structure* model [1];
- We re-train the model in [1] with a *pictorial structure model per class* for better modeling of each class;
- Our proposed model by imposing only *one sub-class (human pose, H) per HOI activity*, examining the importance of allowing a flexible number of pose models to account for the intra-activity variability.

All of the models are trained using the same training data described in Sec.6.1. Following the convention proposed in [11], a body part is considered correctly localized if the endpoints of its segment lie within 50% of the ground-truth segment length from their true positions. Experimental results are shown in Table 1. The percentage correctness tells us that pose estimation still remains a difficult problem. No method offers a solution near 100%. Our full model significantly outperforms the other approaches, even showing a 10% average improvement over a class-based, discrimi-

Table 1. Pose estimation results by our full model and four comparison methods for all testing images. The average part detection percent correctness and standard deviation over 6 HOI classes are presented for each body part. If two numbers are reported in one cell, the left one indicates the left body part and right one indicates the right body part. The best result for each body part is marked in bold font.

Method	Torso	Upper Leg	Lower Leg	Upper Arm	Fore Arm	Head
Iterative parsing [26]	52±19	22±11 22±10	21±9 28±16	24±16 28±17	17±11 14±10	42±18
Pictorial structure [1]	50±14	31±12 30±9	31±15 27±18	18±6 19±9	11±8 11±7	45±8
Class-based pictorial structure	59±9	36±11 26±17	39±9 27±9	30±12 31±12	13±6 18±14	46±11
Our model, only one pose per class	63±5	40±8 36±15	41±10 31±9	38±13 35±10	21±12 23±14	52±8
Our full model	66±6	43±8 39±14	44±10 34±10	44±9 40±13	27±16 29±13	58±11

natively trained pictorial structure model. Furthermore, we can see that allowing multiple poses for each activity class proves to be useful for improving pose estimation accuracy. More sample results are shown in Fig.9, where we visualize the pose estimation results by comparing our model with the state-of-the-art pictorial structure model by [1]. We show that given the object context, poses estimated by our model are less prone to errors that result in strange looking body gestures (e.g. horizontal legs in Fig.9(d)), or a completely wrong location (e.g. nearly all body parts landed on the background in Fig.9(h)).

6.4. Combining object and pose for HOI activity classification

As shown in Fig.5, by inferring the human pose and object in the image, our model gives a prediction of the class label of the human-object interaction. We compare our method with the results reported in [12], and use a bag-of-words representation with a linear SVM classifier as the baseline. The results are shown in Fig.8.

Fig.8 shows that our model significantly outperforms the bag-of-words method and performs slightly better than [12]. Note that the method in [12] uses predominantly the background scene context (e.g. appearance differences in sport courts), which turns out to be highly discriminative among most of these classes of activities. Our method, on the other hand, focuses on the core problem of human-object interactions. It is therefore less data set dependent.

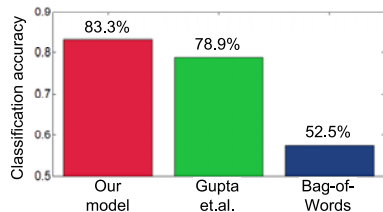


Figure 8. Activity recognition accuracy of different methods: our model, Gupta et al [12], and bag-of-words.

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 random field model that uses a structure learning method to learn important connectivity patterns between objects and human body parts. Experiments show that our model significantly outperforms other state-of-the-art methods in both problems. Our model can be further improved in a number of directions. For example, inspired by [23, 12], we can incorporate useful background scene context to facilitate the recognition of foreground objects and activities. Improving the model to deal with more than one object is also one of the directions of our future research.

Acknowledgement. This research is partially supported by an NSF CAREER grant (IIS-0845230), a Google research award, and a Microsoft Research Fellowship to L.F-F. We also would like to thank Hao Su for helpful comments.

References

- [1] M. Andriluka, S. Roth, and B. Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *CVPR*, 2009. 2, 3, 4, 6, 7, 8
- [2] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE T. Pattern Anal.*, 24(4):509–522, 2002. 3
- [3] I. Biederman, R. Mezzanotte, and J. Rabinowitz. Scene perception: detecting and judging objects undergoing relational violations. *Cognitive Psychol.*, 14:143–177, 1982. 1
- [4] Y. Chen, L. Zhu, C. Lin, A. Yuille, and H. Zhang. Rapid inference on a novel AND/OR graph for object detection, segmentation and parsing. In *NIPS*, 2007. 4, 5, 10
- [5] K. Crammer and Y. Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *J. Mach. Learn. Res.*, 2:265–292, 2001. 5
- [6] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 6
- [7] C. Desai, D. Ramanan, and C. Fowlkes. Discriminative models for multi-class object layout. In *ICCV*, 2009. 1
- [8] S. Divvala, D. Hoiem, J. Hays, A. Efros, and M. Hebert. An empirical study of context in object detection. In *CVPR*, 2009. 1, 2
- [9] M. Everingham, L. V. Gool, C. Williams, J. Winn, and A. Zisserman. The PASCAL VOC2008 Results. 1
- [10] P. Felzenszwalb and D. Huttenlocher. Pictorial structures for object recognition. *Int. J. Comput. Vision*, 61(1):55–79, 2005. 2, 4
- [11] V. Ferrari, M. Marín-Jiménez, and A. Zisserman. Progressive search space reduction for human pose estimation. In *CVPR*, 2008. 6
- [12] A. Gupta, A. Kembhavi, and L. Davis. Observing human-object interactions: Using spatial and functional compatibility for recognition. *IEEE T. Pattern Anal.*, 31(10):1775–1789, 2009. 1, 2, 5, 6, 7



Figure 9. Example testing results of object detection and pose estimation. Each sub-figure contains one testing image, tested on the following four conditions: upper-left→object detection by our model, lower-left→object detection by a scanning window, upper-right→pose estimation by our model, and lower-right→pose estimation by the state-of-the-art pictorial structure method in [1]. Detected objects are shown in double-line red bounding boxes. The color codes for different body parts are: head and torso - magenta, arms - green, legs - cyan. (This figure is best viewed in color.)

- [13] H. Harzallah, F. Jurie, and C. Schmid. Combining efficient object localization and image classification. In *ICCV*, 2009. 1
- [14] G. Heitz and D. Koller. Learning spatial context: using stuff to find things. In *ECCV*, 2008. 1, 2, 4
- [15] J. Henderson. Human gaze control during real-world scene perception. *Trends Cogn. Sci.*, 7(11):498–504, 2003. 1
- [16] M. Hestenes. Multiplier and gradient methods. *J. Optimiz. Theory App.*, 4(5):303–320, 1969. 5, 9
- [17] D. Hoiem, A. Efros, and M. Hebert. Putting objects in perspective. In *CVPR*, 2006. 2
- [18] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009. 4, 9
- [19] C. Lampert, M. Blaschko, and T. Hofmann. Beyond sliding windows: object localization by efficient subwindow search. In *CVPR*, 2008. 2
- [20] B. Leibe, A. Leonardis, and B. Schiele. Combined object categorization and segmentation with an implicit shape model. In *ECCV Workshop on SLCV*, 2004. 1
- [21] J. Liebelt, C. Schmid, and K. Schertler. Viewpoint-independent object class detection using 3D feature maps. In *CVPR*, 2008. 4
- [22] M. Marszalek, I. Laptev, and C. Schmid. Actions in context. In *CVPR*, 2009. 1, 2
- [23] K. Murphy, A. Torralba, and W. Freeman. Using the forest to see the trees: a graphical model relating features, objects, and scenes. In *NIPS*, 2003. 1, 2, 7
- [24] A. Oliva and A. Torralba. The role of context in object recognition. *Trends Cogn. Sci.*, 11(12):520–527, 2007. 1
- [25] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie. Objects in context. In *ICCV*, 2007. 1, 2, 4
- [26] D. Ramanan. Learning to parse images of articulated objects. In *NIPS*, 2006. 2, 3, 4, 6, 7
- [27] X. Ren, A. Berg, and J. Malik. Recovering human body configurations using pairwise constraints between parts. In *ICCV*, 2005. 2
- [28] J. Shotton, J. Winn, C. Rother, and A. Criminisi. TextonBoost: joint appearance, shape and context modeling for multi-class object recognition and segmentation. In *ECCV*, 2006. 1, 2
- [29] B. Taskar, P. Abbeel, and D. Koller. Discriminative probabilistic models for relational data. In *UAI*, 2002. 5
- [30] P. Viola and M. Jones. Robust real-time object detection. *Int. J. Comput. Vision*, 57(2):137–154, 2001. 2, 3
- [31] Y. Wang and G. Mori. Multiple tree models for occlusion and spatial constraints in human pose estimation. In *ECCV*, 2008. 2, 4
- [32] B. Yao and L. Fei-Fei. Grouplet: A structured image representation for recognizing human and object interactions. In *CVPR*, 2010. 1, 2

A. Hill-climbing structure learning

A.1. Objective function

Given a set of images where humans interact with the same class of object (O) with the same type of pose (H), our objective is to learn a connectivity pattern ($\mathcal{C} = \{\mathcal{C}_{OP}, \mathcal{C}_{HP}, \mathcal{C}_{PP}\}$) which best models the interaction between the human and the object. As shown in Fig.3(a), $\mathcal{C}_{OP}$ describes the connection between the object and different body parts, $\mathcal{C}_{HP}$ the connection between the human pose and body parts, and $\mathcal{C}_{PP}$ the connection among different body parts. Note that we learn a connectivity for each pair of human pose and object respectively.

In the learning step, given the locations and size of the object and human body parts, our objective function is

$$\arg \max_{\mathcal{C}} \sum_i \left\{ \sum_{\mathcal{C}_{OP}} \psi_e^i(O, P_n) + \sum_{\mathcal{C}_{HP}} \psi_e^i(H, P_n) + \sum_{\mathcal{C}_{PP}} \psi_e^i(P_m, P_n) + \log \mathcal{N}(|\mathcal{C}|) \right\} \quad (4)$$

where $|\mathcal{C}|$ is the number of edges in $\mathcal{C}$, and $\log(|\mathcal{C}|)$ is a Gaussian prior over the number of edges. $\psi_e^i(\cdot)$ is the potential value computed from the i -th image. Note that in the structure learning stage, we omit the weights of different potential terms. The potential weights will be estimated in the parameter estimation step (Sec.4.2 and Sec.B).

A.2. Hill-climbing structure search with tabu list

We now describe our method of optimizing Eq.4. Note that for each sample i , the value of all the potential terms $\psi_e^i(O, P_n)$, $\psi_e^i(H, P_n)$, and $\psi_e^i(P_m, P_n)$ can be computed by using the Maximum-Likelihood approach. Therefore, given the values of all the potential terms, we use a hill-climbing search [18] method to optimize Eq.4.

In the hill-climbing method, we first randomly initialize the connectivity. Then we execute the following steps repeatedly: We consider all of the solutions that are neighbors of the current one by adding or removing an edge. We compute the score of each solution using Eq.4, from which the one that leads to the best improvement in the score is selected. We continue this process until no improvement can be achieved. Because all the potential terms in Eq.4 can be pre-computed, hill-climbing method converges fast in our problem. The method, however, can only reach a local maximum. There is no guarantee that the local maximum is actually the global optimum. To improve the search result, we adopt the following two approaches.

The first approach is to keep a *tabu list* of operators (adding or deleting a specific edge) that we have recently applied. Then in each search step, we do not consider the operators that reverse the effect of operators applied in the

last five steps. Thus, if we add an edge between two nodes, say the tennis racket and the lower-right-arm, we cannot delete this edge in the next five steps. The tabu list forces the search procedure to explore new directions in the search space so that the performance can be improved [18].

The other approach to reduce the impact of local optimum is *randomization*. We can initialize the connectivity at different starting points, and then use a hill-climbing algorithm for each one, from which the best result is selected. In our method, we use one manually designed starting point and two other random ones. In the manually designed starting point, we connect the human pose node with all the body parts, connect the object with the right-lower-arm, and use a kinematic structure among different body parts.

B. Max-margin parameter estimation

The motivation and formulation of the max-margin learning method has been described in Sec.4.1. Here we elaborate the multiplier method [16] that we use to optimize Eq.3.

We denote $\theta = \{\{\mathbf{w}_r\}_r, \{\xi_i\}_i\}$, $\ell(\theta) = \frac{1}{2} \sum_r \|\mathbf{w}_r\|_2^2 + \beta \sum_i \xi_i$, and use $g_j(\theta)$ to denote all the constraints. Then the training objective function, e.g. Eq.3, becomes finding the minimum value of $\ell(\theta)$ subject to the constraints $\forall j, g_j(\theta) \geq 0$. In order to optimize the problem, we introduce a variable b_j for each $g_j(\theta)$, and convert the original optimization problem to a problem with only equality constraints as the following:

$$\min_{\theta} \ell(\theta), \quad \text{subject to: } \forall j, g_j(\theta) - b_j^2 = 0 \quad (5)$$

Then in the multiplier method, we minimize the following equation,

$$\varphi(\theta, \mathbf{B}, \Gamma, F) = \ell(\theta) - \sum_j \gamma_j (g_j(\theta) - b_j^2) + \frac{F}{2} \sum_j (g_j(\theta) - b_j^2)^2 \quad (6)$$

where $\mathbf{B} = \{b_j\}_j$, F is a suitable large constant, and $\Gamma = \{\gamma_j\}_j$ is the set of Lagrange multipliers. It was shown that when $\ell(\theta)$ is a convex function, we can minimize $\ell(\theta)$ by minimizing $\varphi(\theta, \mathbf{B}, \Gamma, F)$ [16].

In order to minimize $\varphi(\theta, \mathbf{B}, \Gamma, F)$, we first use an analytical method to compute its minimum value with respect to $\mathbf{B}$ and substitute the optimal value $\mathbf{B}^*$ into $\varphi(\theta, \mathbf{B}, \Gamma, F)$, then the problem becomes to optimize $\varphi(\theta, \mathbf{B}^*, \Gamma, F)$. Assuming the optimal Γ^* is known and a suitable large F is given, we can minimize $\varphi(\theta, \mathbf{B}^*, \Gamma^*, F)$ to optimize Eq.5. However the optimal Lagrange multiplier Γ^* is unknown, so we assign it an initial value $\Gamma^{(1)}$ and then iteratively revise it. In the k -th iteration, in order to minimize

$\varphi(\boldsymbol{\theta}, \mathbf{B}^{*(k)}, \Gamma^{(k)}, F)$, we have

$$\frac{\partial \varphi(\boldsymbol{\theta}, \mathbf{B}^{*(k)}, \Gamma^{(k)}, F)}{\partial \boldsymbol{\theta}} = \frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \quad (7)$$

$$- \sum_j \left\{ \left[\gamma_j^{(k)} - F \cdot (g_j(\boldsymbol{\theta}) - b_j^{*(k)})^2 \right] \frac{\partial (g_j(\boldsymbol{\theta}) - b_j^{*(k)})^2}{\partial \boldsymbol{\theta}} \right\}$$

Comparing this equation to the Kuhn-Tucker condition, we revise Γ as the following,

$$\forall j, \quad \gamma_j^{(k+1)} = \gamma_j^{(k)} - F \cdot (g_j(\boldsymbol{\theta}^{(k)}) - b_j^{*(k)})^2 \quad (8)$$

We use the method of completing square to calculate the minimum value of $\varphi(\boldsymbol{\theta}, \mathbf{B}, \Gamma, F)$ with respect to $\mathbf{B}$. $\varphi(\boldsymbol{\theta}, \mathbf{B}, \Gamma, F)$ can be written as

$$\varphi(\boldsymbol{\theta}, \mathbf{B}, \Gamma, F) = Z(\boldsymbol{\theta}) \quad (9)$$

$$+ \sum_j \left\{ \frac{F}{2} \left[b_j^2 - \frac{1}{F} (F g_j(\boldsymbol{\theta}) - \gamma_j) \right]^2 - \frac{\gamma_j^2}{2F} \right\}$$

Then $\min_{\mathbf{B}} \varphi(\boldsymbol{\theta}, \mathbf{B}, \Gamma, F)$ can be obtained when

$$\forall j, \quad b_j^2 = \frac{1}{F} \max \{0, F g_j(\boldsymbol{\theta}) - \gamma_j\} \quad (10)$$

Substituting Eq.10 into Eq.9 and 8, after simplification, we have the following optimization objective

$$\min_{\mathbf{B}} \varphi(\boldsymbol{\theta}, \mathbf{B}, \Gamma, F) = \varphi(\boldsymbol{\theta}, \Gamma, F) \quad (11)$$

$$= \ell(\boldsymbol{\theta}) + \frac{1}{2F} \sum_j \left\{ [\max(0, \gamma_j - F g_j(\boldsymbol{\theta}))]^2 - \gamma_j^2 \right\}$$

and the Lagrange multiplier revision equation

$$\forall j, \quad \gamma_j^{(k+1)} = \max \{0, \gamma_j^{(k)} - F g_j(\boldsymbol{\theta})\} \quad (12)$$

Having obtained Eq.11 and 12, our optimization algorithm is shown in Algorithm 2. We can see that if the algorithm converges very slowly, F will be enlarged.

Given the estimation of $\mathbf{B}^{(k-1)}$, the main component in Algorithm 2 is to find a $\boldsymbol{\theta}$ to minimize $\varphi(\boldsymbol{\theta}, \mathbf{B}^{(k-1)}, \Gamma^{(k)}, F)$. This is an unconstrained minimum problem, we use the steepest descent algorithm to solve this problem, which is shown in Algorithm 3, where $\nabla \varphi(\boldsymbol{\theta}^{(k,l)})$ and $\nabla^2 \varphi(\boldsymbol{\theta}^{(k,l)})$ denote the first and second order derivations of $\boldsymbol{\theta}^{(k,l)}$ respectively. In the implementation we set $\alpha = 2$, $\beta = 0.8$, $F = 50$.

C. Compositional inference

In the inference process, given a model and an image, we need to find the optimal spatial configurations of the object and the body parts of the human. We use the compositional inference method described in [4] to achieve this

Input: Initial $\boldsymbol{\theta}^{(0)}$, initial estimation of $\mathbf{B}^{(0)}$, $\Gamma^{(1)}$, and F , constants $\alpha > 1$, $\beta \in (0, 1)$, error tolerance value $\varepsilon > 0$. Let $k = 1$.

*: Use $\boldsymbol{\theta}^{(k-1)}$ as the initial point, solve $\boldsymbol{\theta}^{(k)} = \min_{\boldsymbol{\theta}} \varphi(\boldsymbol{\theta}, \mathbf{B}^{(k-1)}, \Gamma^{(k)}, F)$;

Use Eq.10 to compute $\mathbf{B}^{(k)}$;

if $\|g_j(\boldsymbol{\theta}^{(k)}) - b_j^{(k)}\| < \varepsilon$ **then** Stop the algorithm;

if $\frac{\|g_j(\boldsymbol{\theta}^{(k)}) - b_j^{(k)}\|^2}{\|g_j(\boldsymbol{\theta}^{(k-1)}) - b_j^{(k-1)}\|^2} \geq \beta$ **then** Let $F = \alpha F$;

Use Eq.12 to get Γ_j ;

Let $k = k + 1$, return to *;

Output: The parameters $\boldsymbol{\theta}^{(k)}$.

Algorithm 2: The multiplier and gradient method.

Set the initial value $\boldsymbol{\theta}^{(k,1)}$, error tolerance $\varepsilon' > 0$. Let $l = 1$;

*: Determine the optimal search direction

$$d^{(l)} = -\nabla \varphi(\boldsymbol{\theta}^{(k,l)});$$

if $\|d^{(l)}\| < \varepsilon'$ **then** Stop the algorithm;

Calculate the best step value λ_l in the direction of $d^{(l)}$,

$$\lambda_l = \frac{\nabla \varphi(\boldsymbol{\theta}^{(k,l)})^T \cdot \nabla \varphi(\boldsymbol{\theta}^{(k,l)})}{\nabla \varphi(\boldsymbol{\theta}^{(k,l)})^T \cdot \nabla^2 \varphi(\boldsymbol{\theta}^{(k,l)}) \cdot \nabla \varphi(\boldsymbol{\theta}^{(k,l)})};$$

Compute $\boldsymbol{\theta}^{(k,l+1)} = \boldsymbol{\theta}^{(k,l)} - \lambda_l \nabla \varphi(\boldsymbol{\theta}^{(k,l)})$;

Let $l = l + 1$, go to *;

Output: $\boldsymbol{\theta}^{(k,l)}$.

Algorithm 3: The steepest descent algorithm.

goal. The algorithm has a bottom-up stage which makes proposals of different parts. The bottom-up stage starts from the object detection and human body parts detection scores, from which we obtain the image parts with large detection scores for further processing. Then in the first level of the bottom-up stage, if two nodes of the body parts or the object are connected, we enumerate all combinations of the strong detection responses of the two nodes. The combinations with low fitness scores are removed. We compute the fitness score by adding the detection scores of the two parts, as well as the potential value of the edge between them. A clustering method is applied to the remaining combinations to obtain a small set of max-proposals. Then the remaining proposals are merged according to the connectivity structure among different image parts. In the compositional inference stage, we omit the weights of different potentials and set all of them to 1. Please refer to [4] for more details about this inference method. In the first level of the bottom-up stage, we propose 5000 node combinations for each edge. After clustering only 30~100 combinations are remained.