Pairwise Body-Part Attention for Recognizing Human-Object Interactions

Hao-Shu Fang[1][0000−0002−0758−0293], Jinkun Cao[1], Yu-Wing Tai[2], and Cewu Lu[1]*[0000−0002−4023−9257]

[1] Shanghai Jiao Tong University, China
fhaoshu@gmail.com, {caojinkun,lucewu}@sjtu.edu.cn
[2] Tencent YouTu Lab, China
yuwingtai@tencent.com

Abstract. In human-object interactions (HOI) recognition, conventional methods consider the human body as a whole and pay a uniform attention to the entire body region. They ignore the fact that normally, human interacts with an object by using some parts of the body. In this paper, we argue that different body parts should be paid with different attention in HOI recognition, and the correlations between different body parts should be further considered. This is because our body parts always work collaboratively. We propose a new pairwise body-part attention model which can learn to focus on crucial parts, and their correlations for HOI recognition. A novel attention based feature selection method and a feature representation scheme that can capture pairwise correlations between body parts are introduced in the model. Our proposed approach achieved **10**% relative improvement (36.1 mAP→ 39.9 mAP) over the state-of-the-art results in HOI recognition on the HICO dataset. We will make our model and source codes **publicly available**.

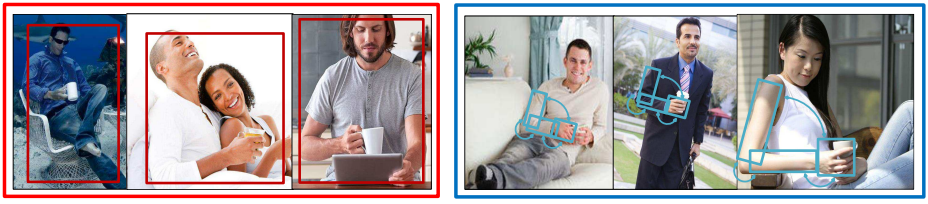
Keywords: Human-Object Interactions, Body-Part Correlations, Attention Model

1 Introduction

Recognizing Human-Object Interactions (HOI) in a still image is an important research problem and has applications in image understanding and robotics [1, 44, 48]. From a still image, HOI recognition needs to infer the possible interactions between the detected human and objects. Our goal is to evaluate the probabilities of certain interactions on a predefined HOI list.

Conventional methods consider the problem of HOI recognition at holistic body level [40, 21, 52] or very coarse part level (e.g., head, torso, and legs) [11] only. However, studies in cognitive science [35, 4] have already found that our visual attention is non-uniform, and humans tend to focus on different body

* The corresponding author is Cewu Lu, email: lucewu@sjtu.edu.cn, twitter: @Cewu_Lu, Cewu Lu is a member of MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, and SJTU-SenseTime AI lab.



(a) Conventional HOI recognition model (b) Our model

Fig. 1. Given an image, a person holding a mug in his/her hand, conventional model (a) infers the HOI from the whole body feature. In contrast, our model (b) explicitly focuses on discriminative body parts and the correlations between objects and different body parts. In this example, the upper and lower arms which hold a mug form an acute angle across all of the above images.

parts according to different context. As shown in Figure 1, although the HOI label are the same across all examples, the body gestures are all different except for the arm which holds a mug. This motivates us to introduce a non-uniform attention model which can effectively discover the most informative body parts for HOI recognition.

However, simply building attention on body parts can not capture important HOI semantics, since it ignores the correlations between different body parts. In Figure 1, the upper and lower arms and the hand work collaboratively and form an acute angle due to physical constraints. Such observation motivates us to further focus on the correlations between multiple body parts. In order to make a practical solution, we consider the joint correlations between each pair of body parts. Such pairwise sets define a new set of correlation feature maps whose features should be extracted simultaneously. Specifically, we introduce pairwise ROI pooling which pools out the joint feature maps of pairwise body parts, and discards the features of other body parts. This representation is robust to irrelevant human gestures and the detected HOI labels have significantly less false positives, since the irrelevant body parts are filtered. With the set of pairwise features, we build an attention model to automatically discover discriminative pairwise correlations of body parts that are meaningful with respect to each HOI label. By minimizing the end-to-end loss, the system is forced to select the most representative pairwise features. In this way, our trained pairwise attention module is able to extract meaningful connections between different body parts.

To the best of our knowledge, our work is the first attempt to apply the attention mechanism to human body part correlations for recognizing human-object interactions.

We evaluate our model on the HICO dataset [5] and the MPII dataset [2]. Our method achieves the state-of-the-art result, and outperforms the previous methods by **10%** relatively in mAP on HICO dataset.

2 Related work

Our work is related to two active areas in computer vision: human-object interactions and visual attention.

Human-Object Interactions Human-object interactions (HOI) recognition is a sub-task of human actions recognition but also a crucial task in understanding the actual human action. It can resolve the ambiguities in action recognition when two persons have almost identical pose and provide a higher level of semantic meaning in the recognition label. Early researches in action recognition consider video inputs. Representative works include [16, 41, 42]. In action recognition from still images, previous works attempt to use human pose to recognize human action [43, 40, 21, 47, 28, 52].

However, considering human pose solely is ambiguous since there is no motion cue in a still image. Human-object interactions are introduced in order to resolve such ambiguities. With additional high level contextual information, it has demonstrated success in improving performance of action recognition [8, 51, 32, 20]. Since recognizing the small object is difficult, some works [50, 54, 36] attempt to ease the object recognition by recognizing discriminative image patches. Other lines of work include utilizing high level attributes in images [26, 53], exploring the effectiveness of BoF method [6], incorporating color information [24] and semantic hierarchy [33] to assist HOI recognition.

Recently, deep learning based methods [12, 11, 29, 13] give promising results on this task. Specifically, Gkioxari *et al.* [11] develop a part based model to make fine-grained action recognition based on the input of both whole-person and part bounding boxes. Mallya and Lazebnik [29] propose a simple network that fuses features from a person bounding box and the whole image to recognize HOIs.

Comparing to the aforementioned methods, especially the deep learning based methods, our method differs mainly in the following aspects. Firstly, our method explicitly considers human body parts and their pairwise correlations, while Gkioxari *et al.* [11] only consider parts at a coarse level (i.e., head, torso and legs) and the correlations among them are ignored, and Mallya *et al.* [29] only consider bounding boxes of the whole person. Secondly, we propose an attention mechanism to learn to focus on specific parts of body and the spatial configurations, which has not been discussed yet in the previous literatures.

Attention model Human perception focuses on parts of the field of view to acquire detailed information and ignore those irrelevant. Such attention mechanism has been studied for a long time in computer vision community. Early works motivated by human perception are saliency detection [22, 19, 15]. Recently, there have been works that try to incorporate attention mechanism into deep learning framework [31, 25, 7]. Such attempt has been proved to be very effective in many vision tasks including classification [45], detection [3], image captioning [55, 38, 46] and image-question-answering [49]. Sharma *et al.* [37] first applied attention model to the area of action recognition by using LSTM [18] to focus on important parts of video frames. Several recent works [27, 39, 10] are partly related to

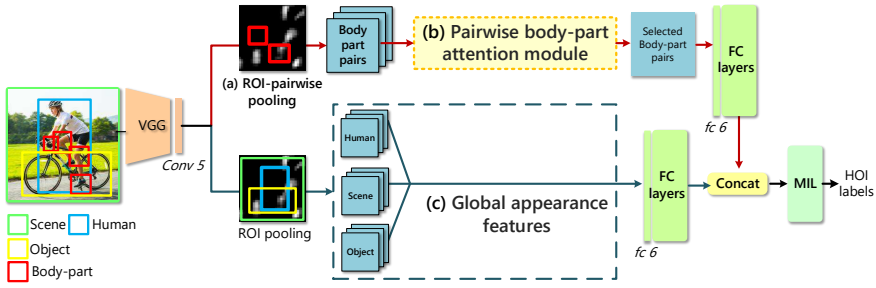


Fig. 2. Overview of our framework. The model first extracts visual features of human,object and scene from a set of proposals. We encode the features of different body parts and their pairwise correlations using ROI-pairwise pooling (a). Then our pairwise body-part attention module (b) will select the feature maps of those discriminative body-part pairs. The global appearance features (c) from the human, object and scene will also contribute to the final predictions. Following [29], we adopt MIL to address the problem of multi-person co-occurrence in an image. See text for more details.

our paper. In [27, 39], a LSTM network is used to learn to focus on informative joints of skeleton within each frame to recognize actions in videos. Their method differs from ours that their model learns to focus on discriminative joints of 3D skeleton in an action sequence. In [10], the authors introduce an attention pooling mechanism for action recognition. But their attention is applied to the whole image instead of explicitly focusing on human body parts and the correlations among body parts as we do.

3 Our Method

Our approach utilizes both global and local information to infer the HOI labels.

The global contextual information has been well studied by many previous works [8, 51, 32, 20], focusing on utilizing the features of person, object and scene. In section 3.1, we review the previous deep learning model [29] that utilizes features of person and scene. Based on the model from [29], we further incorporate object features. This forms a powerful base network which efficiently captures global information. Note that our improved base network has already achieved better performance than the model presented by [29].

In section 3.2, we describe our main algorithm to incorporate pairwise body parts correlations into the deep neural network. Specifically, we propose a simple yet efficient pooling method called ROI-pairwise pooling which encodes both local features of each body part and the pairwise correlations between them. An attention model is developed to focus on discriminative pairwise features. Finally, we present the combination of global features and our local pairwise correlation features in section 3.3. Figure 2 shows an overview of our network architecture.

3.1 Global Appearance Features

Scene and Human Features To utilize the features of the whole person and the scene for HOI recognition, [29] proposed an effective model and we adopt it to build our base network. As shown in Fig. 2, given an input image, we resized and forwarded it through the VGG convolutional layers until the *Conv5* layer. On this shared feature maps, the ROI pooling layer extracts ROI features for each person and the scene given their bounding boxes. For each detected person, the features of him/her are concatenated with the scene features and forwarded through fully connected layers to estimate the scores of each HOI on the predefined list. In the HICO dataset, there can be multiple persons in the same image. Each HOI label is marked as positive as long as the corresponding HOI is observed. To address the issue of multiple persons, the *Multiple Instance Learning*(MIL) framework [30] is adopted. The inputs of MIL layer are the predictions for each person in the image, and the output of it is a score array which takes the maximum score of each HOI among all the input predictions. Since MIL is not the major contribution of our work, we refer readers to [29, 30] for more details of MIL and how it is applied in HOI recognition.

Incorporating Object Features In order to have a coherent understanding of the HOI in context, we further improve the baseline method by incorporating object features, which is ignored in [29].

Feature Representation Given an object bounding box, a simple solution is to extract the corresponding feature maps and then concatenate them with the existing features of human and scene. However, such method does not have much improvement for the task of HOI recognition. This is because the relative locations between object and human are not encoded. So instead, we set our ROI as a union box of detected human and object. Our experiments (Section 4.2) show that such representation is effective.

union box of person and object

Handling Multiple Objects In HICO dataset, there can be multiple persons and multiple objects in an image. For each person, multiple objects can co-appear around him/her. To solve this problem, we sample multiple union boxes of different objects and the person, and the ROI pooling is applied to each union box respectively. The total number of sampled objects around a person is fixed in our implementation. Implementing details will be explained in Sec. 4.

The extracted features of objects are concatenated together with the features of human and scene. This builds a strong base network for capturing well global appearance features.

3.2 Local Pairwise Body-part Features

In this subsection, we will describe how to obtain pairwise body-part features using our pairwise body-part attention module.

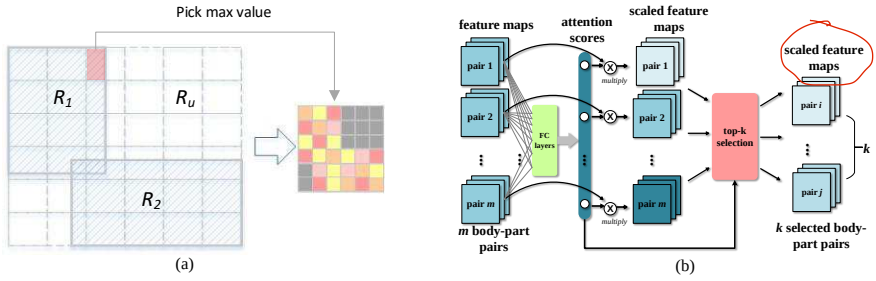


Fig. 3. (a) Illustration of the ROI-pairwise pooling layer. The R_1 and R_2 each represent a bounding box of different body parts. The ROI-pairwise pooling layer extracts the union area feature of R_1 and R_2. The remaining areas are discarded. For each sampled grid location in the ROI-pairwise pooling, the maximum value within the grid area is sampled. (b) Pipeline of the pairwise body-part attention module. From the pairwise body part feature maps pooled by the ROI-pairwise pooling layer, we apply FC layers to estimate the attention score. The attention score is then multiplied with the body part feature maps. Finally, we introduce the feature selection layer which selects the top k most important body part pairs and their scaled feature maps are propagated to the next step.

ROI-pairwise Pooling Given a pair of body parts, we want to extract their joint feature maps while preserving their relative spatial relationships. Let us denote the ROI pair by $R_1(r_1, c_1, h_1, w_1)$, $R_2(r_2, c_2, h_2, w_2)$, and their union box by $R_u(r_u, c_u, h_u, w_u)$, where (r, c) specifies the top-left corner of the ROI and (h, w) specifies the height and width. An intuitive idea is to set the ROI as the union box of the body-part pair and use ROI pooling layer to extract the features. However, when the two body parts are far from each other, e.g., the wrist and the ankle, their union box would cover a large area of irrelevant body-part. These irrelevant features will confuse the model during training. To avoid it, we assign activation outside (two) body-part boxes as zero to eliminate those irrelevant features. Then, to ensure the uniform size of R_u representation, we convert the feature map of union box R_u into a fixed size of $H \times W$ feature. It works in a uniformly max-pooling manner: we first divide the $h_u \times w_u$ into $H \times W$ grids, then for each grid, the maximum value inside that grid cell is pooled into the corresponding output cell. Figure 3(a) illustrates the operation of our ROI-pairwise pooling.

With ROI-pairwise pooling layer, both the joint features of two body parts and their relative location are encoded. Note that the number of body-part pairs are usually big ($C(n, 2)$ for n parts) and many pairwise body parts are rarely correlated. We automatically discover those discriminative correlations by proposing an attention module.

Attention Module Figure 3 (b) illustrates the pipeline of our attention module. Our attention module takes the feature maps of all possible pairwise body-

part pairs $P = \{p_1, p_2, ..., p_m\}$ after the ROI-pairwise pooling as input, where $m = C(n, 2)$ is the number of body-part pairs. For each pairwise body-part p_i, the fully connected layer would regress an attention score s_i. The scores $S = \{s_1, s_2, ..., s_m\}$ for m pairwise body-parts indicate the importance of each body-part pair.

Feature Selection As aforementioned, only some body part pairs are relevant to HOI and irrelevant ones may cause over-fitting of neural network. Assuming that we need to select features of k body-part pairs, our selection layer will keep the feature maps that belong to the body-part pairs with top-k score and drop the remaining. The selected set can be expressed as:

$$\Phi = \{p_i | s_i \text{ ranks top } k \text{ in } S\}. \tag{1}$$

Attention Allocation Different feature maps always have equal value scale, yet they offer different contributions on HOI recognition. So, we should re-scale the feature maps to reflect their indeed influence. Mathematically, it is modeled as multiplying the corresponding attention score, which can be expressed as:

$$f_j = p_{c(j)} \times s_{c(j)}, \tag{2}$$

where $c(j)$ is the index for the j^{th} element in Φ and f_j represents the j^{th} re-scaled feature maps.

Discussion We only allow k pairwise features to represent an interaction. S is forced to assign large value to some pairwise body parts related with input interaction to achieve better accuracy. Therefore, S enables attention mechanism without human supervision. In the experiment section 4.4, we verify that the learned attention score is in accord with human perception.

Training Since Eqn. (1) is not a differentiable function, it has no parameter to be updated and only conveys gradients from the latter layer to the former one during back-propagation. When only the top k pairwise feature maps are selected, the gradients of the feature maps that are selected by the feature selection layer will be copied from latter layer to the former layer. The gradients of the dropped feature maps will be discarded by setting the corresponding values to zero. Since Eqn.(2) can be derived easily, the attention scores are updated automatically during back-propagation and our attention module is trained in an end-to-end manner.

Combining the ROI-pairwise pooling layer and the attention module, our pairwise body-part attention module has the following properties:

- Both local features of each body part and the higher level spatial relationships between body parts are taken into consideration.
- For different HOI, our novel pairwise body-part attention module will automatically discover the discriminative body parts and pairwise relationships.

3.3 Combining Global and Local Features

After obtaining the selected pairwise body-part features and the global appearance features, we forwarded them through the last FC layers respectively to estimate the final predictions. The prediction is applied for every detected person instances.

4 Experiment

We report our experimental results in this section. We first describe the experimental setting and the details in training our baseline model. Then, we compare our results with those of state-of-the-art methods. Ablation studies are carried to further analyze the effectiveness of each component of our network. Finally, some analyses will be given at the end of this section.

4.1 Setting

Dataset We conduct experiments on two frequently used datasets, namely, HICO and MPII dataset. **HICO dataset** [5] is currently the largest dataset for HOI recognition. It contains 600 HOI labels in total and multiple labels can be simultaneously presented in an image. The ground truth labels are given at image level without any bounding box or location information. Also, multiple persons can appear in the same image, and the activities they perform may or may not be the same. Thus the label can be regarded as an aggregation over all HOI activities in an image. The training set contains 38,116 images and the testing set contains 9,658 images. We randomly sample 10,000 images from the training set as our validation set. **MPII dataset** [2] contains 15,205 training images and 5708 test images. Unlike HICO dataset, all person instances in an image are assumed to take the same action and each image is classified into only one of 393 action classes. Following [29], we sample 6,987 images from the training set as validation set.

HICO We use Faster RCNN [34] detector to obtain human and object bounding boxes. For each image, 3 human proposals and 4 object proposals will be sampled to fit the GPU memory. If the number of human or objects is less than expected, we pad the remaining area with zero. For the human body parts, we first use pose estimator [9] to detect all human keypoints and then define 10 body parts based on keypoints. The selected representative human body parts of our method are shown in Figure 5 (a). Each part is defined as a regular bounding box with side length proportional to the size of detected human torso. For body-part pairs, the total number of the pair-wise combination between different body parts is $45(C(10, 2))$.

We first try to reproduce Mallya & Lazebnik [29]'s result as our baseline. However, with the best of our effort, we can only achieve 35.6 mAP, while the

reported result from Mallya and Lazebnik is 36.1 mAP. We use this model as our baseline model. During training, we follow the same setting as [29], with an initial learning rate of 1e-5 for 30000 iterations and then 1e-6 for another 30000 iterations. The batch size is set to 10. Similar to the work in [29, 14], the network is fine-tuned until *conv3* layer. We train our model using Caffe framework [23] on a single Nvidia 1080 GPU. In the testing period, one forward pass takes 0.15s for an image.

Since the HOI labels in the HICO dataset are highly imbalanced, we adopt a weighted sigmoid cross entropy loss

$$\text{loss}(I, y) = \sum_{i=1}^{C} w_p^i \cdot y^i \cdot \log(\hat{y}^i) + w_n^i \cdot (1 - y^i) \cdot \log(1 - \hat{y}^i),$$

where C is the number of independent classes, w_p and w_n are weight factors for positive and negative examples, $\hat{y}$ is model's prediction and y is the label for image I. Following [29], we set $w_p = 10$ and $w_n = 1$.

MPII Since all persons in an image are performing the same action, we directly train our model on each person instead of using MIL. The training set of MPII contains manually labeled human keypoints. For testing set, we ran [9] to get human keypoints and proposals. The detector [34] is adopted to obtain object bounding boxes in both training and testing sets. Similar to the setting for HICO dataset, we sample a maximum of 4 object proposals per image. During training, we set our initial learning rate as 1e-4, with a decay of 0.1 for every 12000 iterations and stop at 40000 iterations. For MPII dataset, we do not use the weighted loss function for fair comparison with [29].

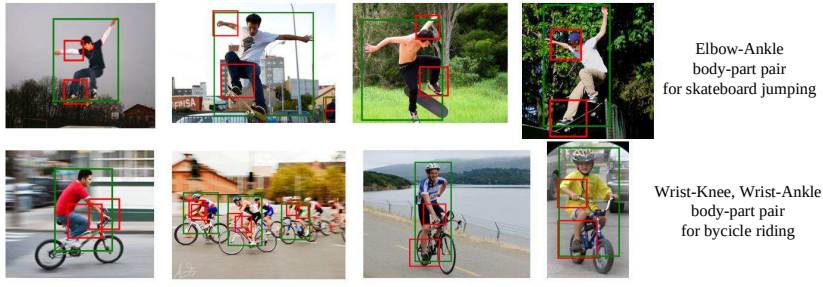
4.2 Results

Method	Full Im.	Bbox/Pose	MIL	Wtd Loss	mAP
AlexNet+SVM [5]	✓				19.4
R*CNN [14]		✓	✓		28.5
Mallya & Lazebnik [29]	✓	✓	✓		33.8
Pose Regu. Attn. Pooling [10]	✓	✓			34.6
Ours	✓	✓	✓		**37.5**
Mallya & Lazebnik, weighted loss [29]	✓	✓	✓	✓	36.1
Ours, weighted loss	✓	✓	✓	✓	**39.9**

Table 1. Comparison with previous results on the HICO test set. The result of R*CNN is directly copied from [29].

Method	Full Img	Bbox	Pose	Val (mAP)	Test (mAP)
Dense Trajectory + Pose [2]	✓		✓	-	5.5
R*CNN, VGG16 [14]		✓		21.7	26.7
Mallya & Lazebnik, VGG16 [29]	✓	✓		-	32.2
Ours, VGG16	✓	✓	✓	**30.9**	**36.8**
Pose Reg. Attn. Pooling, Res101 [10]	✓		✓	30.6	36.1
Ours, Res101	✓	✓	✓	**32.0**	**37.5**

Table 2. Comparison with previous results on the MPII test set. The results on test set are obtained by e-mailing our predictions to the author of [2]



(a) Our model is able to discover correlations between different body-parts and tends to pick similar body-part pairs for each HOI. The body part pairs with the highest attention score are shown in the red boxes.

(b) Some examples of our model's predictions. The first two rows are results from HICO dataset and the last row is results from MPII dataset. The detected human bounding boxes are shown in the green boxes and the body part pairs with the highest attention score are shown in the red boxes. Predicted HOIs are given underneath.

Fig. 4. Results of our model's predictions.

We compare our performance on HICO testing set in Table 1 and on MPII testing set in Table 2. By selectively focusing on human body parts and their correlations, our VGG16 based model achieves **37.6** mAP on HICO testing set and **36.8** mAP on MPII testing set. Using a weighted loss function, we can further achieve **39.9** mAP on HICO testing set. Since [10] use ResNet101 [17] as their base model, we also perform an experiment on MPII dataset by replacing our VGG16 base network with the ResNet101 for fair comparison with [10]. We can see that our VGG16 based model has already achieved better performance than [10] on HICO and MPII dataset, and by using the same base model, we outperform [10] by 1.4 mAP on MPII dataset. These results show that the information from body-parts and their correlations is important in recognizing human-object interactions, and it allows us to achieve the state-of-the-art performances on both datasets.

Figure 4 shows some qualitative results produced by our model. We visualize the body-part pairs with the highest attention score in the red boxes. More results are given in supplementary material.

4.3 Ablative studies

To evaluate the effectiveness of each component in our network, we conduct several experiments on HICO dataset and the results are shown in Table 3.

	Method	mAP
a)	baseline	35.6
b)	union box	**37.0**
	tight box	36.3
c)	body parts, w/o attention	38.0
d)	body-part pairs, w/o attention	38.9
	body-part pairs, with attention	**39.9**
	body parts & pairs, with attention	39.1

Table 3. Performance of the baseline networks on the HICO test set. "union box" refers to the features of an object which are extracted from the area of union box of human and object. "tight box" refers to the features of an object which are extracted from the exact area of the object tight box. "w/o attention" refers to the method without attention mechanism.

Incorporating Object Information As shown in Table 3(b), our improved baseline model with object features can achieve higher mAP than the baseline method without using object features. It shows that object information is important for HOI recognition. From the table, we can see that using the features from the union box instead of the tight box can achieve higher mAP. Note that our improved baseline model has already achieved the state-of-the-art results with 0.9 mAP higher than the results reported by [29].

Improvements from body parts information We evaluate the performance improvement with additional body-parts information. The feature maps of 10 body parts are directly concatenated with the global appearance features, without taking the advantages of attention mechanism or body-part correlations. As can be seen in Table 3(c), we further gain an improvement of 1.0 mAP.

Pairwise Body-part Attention To evaluate the effectiveness of each component of our pairwise body-part attention model, a series of experiments have been carried out and results are reported in Table 3(d).

Firstly, we consider the correlations of different body parts. The feature maps of the 45 body-part pairs are concatenated with the global appearance features to estimate HOI labels. With body-part pairwise information considered, our model can achieve **38.9** mAP. It demonstrates that exploiting spatial relationships between body parts benefits the task of HOI recognition.

Then, we add our attention module upon this network. For our feature selection layer, we set k as 20. The influence of the value of k will be discussed in the analysis in 4.4. With our pairwise body-part attention model, the performance of our model further yields **39.9** mAP even though the fully connected layers receive less information from fewer parts.

We also conduct an experiment by simultaneously learning to focus on discriminative body parts and body-part pairs. The candidates for our attention model are the feature maps of 10 body parts and 45 body-part pairs. However, the final result drops slightly to 39.1 mAP. One possible reason is that our ROI-pairwise pooling has already encoded local features of each single body part. The extra information of body parts may have distracted our attention network.

4.4 Analysis

Parameter for Feature Selection Layer In our feature selection layer, we need to decide k, the number of body part pairs that we propagate to the next step. We perform an experiment to evaluate the effect of k. We train our pairwise body part attention model on HICO training set with different value of k. The performances on validation set are reported in Figure 5 (b). When k increases, the performance of our model increases until $k = 20$. After that, the performance of our model starts to drop. When k equals to 45, it is equivalent to not using the feature selection layer. The performance in this case is 1.2 mAP lower than the highest accuracy. This indicates that rejecting irrelevant body-part pairs is important.

Evaluation of Attention To see how close the attention of our model is to human's attention, we list out different HOIs and their corresponding body-part pairs that are selected most frequently by our trained attention module. Some examples are presented in Table 4. The entire list is provided in supplementary material. We invite 30 persons to judge whether the choice of the selected pairs are relevant to the given HOI labels. If half of the persons agree that a selected

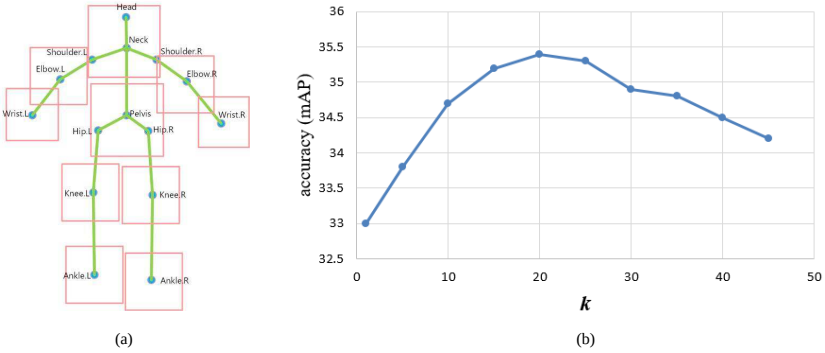


Fig. 5. (a) Our defined human body-parts. Each bounding box denotes a defined body part. (b) The relationship between recognition accuracy and the number of selected pairwise body part feature maps in the feature selection layer.

HOI	Selected correlations		
chase-bird	l.knee-r.wrist	r.elbow-neck	r.ankle-r.elbow
board-car	r.ankle-l.elbow	r.ankle-r.elbow	r.elbow-neck
hug-person	l.elbow-neck	r.elbow-neck	r.wrist-neck
jump-bicycle	l.wrist-pelvis	r.ankle-pelvis	r.elbow-neck
adjust-tie	r.wrist-neck	l.wrist-neck	l.elbow-neck

Table 4. Some HOIs and their corresponding most selected body-part pairs chosen by our model. The "l" and "r" flags denote for left and right.

body part pair is important to decide the HOI labels, we regard the selected body part pair as correct. In our setting, the top-k accuracy means the correct body part pair appears in the first k predictions of attention module. Our top-1 accuracy achieves 0.28 and top-5 accuracy achieves 0.76. It is interesting to see that the body part pairs selected by our attention module match with our intuition to some extent.

Improvements by HOI class To see which kinds of interactions become less confused due to the incorporation of body part information, we compare the results on 20 randomly picked HOIs in HICO dataset with and without the proposed pairwise body-part attention module. The comparisons are summarized in Table 5. When the HOIs require more detailed body part information, such as surfboard holding, apple buying and bird releasing, our model shows a great improvement over the baseline model.

HOI	[29]	Ours	HOI	[29]	Ours
cat scratching	47.7	**50.9**	train boarding	37.1	**48.2**
umbrella carrying	83.7	**86.9**	apple buying	19.3	**59.0**
keyboard typing on	**71.6**	68.3	cake lighting	16.3	**24.1**
boat inspecting	21.1	**31.9**	cup inspecting	1.0	**1.5**
oven cleaning	**22.1**	13.1	fork licking	4.4	**5.3**
surfboard holding	52.9	**63.6**	bird releasing	14.5	**51.3**
dining table eating at	86.6	**86.9**	car parking	**28.9**	26.3
sandwich no interaction	74.2	**85.2**	horse jumping	**87.0**	86.9
motorcycle washing	57.7	**64.8**	spoon washing	14.5	**15.3**
airplane loading	**64.1**	60.0	toilet repairing	11.4	**22.6**

Table 5. We randomly pick 20 categories in HICO dataset and compare our results with results from Mallya&Lazebnik [29]. The evaluation metric is mAP. The full set of results can be found in the supplementary materials.

5 Conclusions

In this paper, we have proposed a novel pairwise body part attention model which can assign different attention to different body-part pairs. To achieve our goal, we have introduced the ROI pairwise pooling, and the pairwise body-part attention module which extracts useful body part pairs. The pairwise feature maps selected by our attention module are concatenated with background, human, and object features to make the final HOI prediction. Our experimental results show that our approach is robust, and it significantly improves the recognition accuracy especially for the HOI labels which require detailed body part information. In the future, we shall investigate the possibility of including multi-person interactions into the HOI recognition.

6 Acknowledgement

This work is supported in part by the National Key R&D Program of China No. 2017YFA0700800, National Natural Science Foundation of China under Grants 61772332 and SenseTime Ltd.

References

1. Aksoy, E.E., Abramov, A., Dörr, J., Ning, K., Dellen, B., Wörgötter, F.: Learning the semantics of object–action relations by observation. The International Journal of Robotics Research **30**(10), 1229–1249 (2011)
2. Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: 2d human pose estimation: New benchmark and state of the art analysis. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2014)
3. Ba, J., Mnih, V., Kavukcuoglu, K.: Multiple object recognition with visual attention. In: arXiv preprint arXiv:1412.7755 (2014)
4. Boyer, T.W., Maouene, J., Sethuraman, N.: Attention to body-parts varies with visual preference and verb–effector associations. Cognitive Processing (2017)
5. Chao, Y.W., Wang, Z., He, Y., Wang, J., Deng, J.: Hico: A benchmark for recognizing human-object interactions in images. In: ICCV (2015)
6. Delaitre, V., Laptev, I., Sivic, J.: Recognizing human actions in still images: a study of bag-of-features and part-based representations. In: BMVC (2010)
7. Denil, M., Bazzani, L., Larochelle, H., de Freitas, N.: Learning where to attend with deep architectures for image tracking. Neural computation **24**(8), 2151–2184 (2012)
8. Desai, C., Ramanan, D., Fowlkes, C.: Discriminative models for static human-object interactions. In: CVPR'w (2010)
9. Fang, H.S., Xie, S., Tai, Y.W., Lu, C.: RMPE: Regional multi-person pose estimation. In: ICCV (2017)
10. Girdhar, R., Ramanan, D.: Attentional pooling for action recognition. In: NIPS (2017)
11. Gkioxari, G., Girshick, R., Malik, J.: Actions and attributes from wholes and parts. In: ICCV (2015)
12. Gkioxari, G., Hariharan, B., Girshick, R., Malik, J.: R-cnns for pose estimation and action detection. In: arXiv preprint arXiv:1406.5212 (2014)
13. Gkioxari, G., Girshick, R., Dollár, P., He, K.: Detecting and recognizing human-object intaractions. In: arXiv preprint arXiv:1704.07333 (2017)
14. Gkioxari, G., Girshick, R., Malik, J.: Contextual action recognition with r* cnn. In: ICCV (2015)
15. Goferman, S., Zelnik-Manor, L., Tal, A.: Context-aware saliency detection. TPAMI **34**(10), 1915–1926 (2012)
16. Han, D., Bo, L., Sminchisescu, C.: Selection and context for action recognition. In: ICCV (2009)
17. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. arXiv preprint arXiv:1512.03385 (2015)
18. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
19. Hou, X., Zhang, L.: Saliency detection: A spectral residual approach. In: CVPR (2007)
20. Hu, J.F., Zheng, W.S., Lai, J., Gong, S., Xiang, T.: Recognising human-object interaction via exemplar based modelling. In: ICCV (2013)
21. Ikizler, N., Cinbis, R.G., Pehlivan, S., Duygulu, P.: Recognizing actions from still images. In: ICPR (2008)
22. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. TPAMI **20**(11), 1254–1259 (1998)

23. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. arXiv preprint arXiv:1408.5093 (2014)
24. Khan, F.S., Anwer, R.M., van de Weijer, J., Bagdanov, A.D., Lopez, A.M., Felsberg, M.: Coloring action recognition in still images. IJCV **105**(3), 205–221 (2013)
25. Larochelle, H., Hinton, G.E.: Learning to combine foveal glimpses with a third-order boltzmann machine. In: NIPS (2010)
26. Liu, J., Kuipers, B., Savarese, S.: Recognizing human actions by attributes. In: CVPR (2011)
27. Liu, J., Wang, G., Hu, P., Duan, L.Y., Kot, A.C.: Global context-aware attention lstm networks for 3d action recognition. In: CVPR (2017)
28. Maji, S., Bourdev, L., Malik, J.: Action recognition from a distributed representation of pose and appearance. In: CVPR (2011)
29. Mallya, A., Lazebnik, S.: Learning models for actions and person-object interactions with transfer to question answering. In: ECCV (2016)
30. Maron, O., Lozano-Pérez, T.: A framework for multiple-instance learning (1998)
31. Mnih, V., Heess, N., Graves, A., et al.: Recurrent models of visual attention. In: NIPS (2014)
32. Prest, A., Schmid, C., Ferrari, V.: Weakly supervised learning of interactions between humans and objects. TPAMI **34**(3), 601–614 (2012)
33. Ramanathan, V., Li, C., Deng, J., Han, W., Li, Z., Gu, K., Song, Y., Bengio, S., Rosenberg, C., Fei-Fei, L.: Learning semantic relationships for better action retrieval in images. In: CVPR (2015)
34. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: NIPS (2015)
35. Ro, T., Friggel, A., Lavie, N.: Attentional biases for faces and body parts. Visual Cognition **15**(3), 322–348 (2007)
36. Sharma, G., Jurie, F., Schmid, C.: Expanded parts model for human attribute and action recognition in still images. In: CVPR (2013)
37. Sharma, S., Kiros, R., Salakhutdinov, R.: Action recognition using visual attention (2015)
38. Shih, K.J., Singh, S., Hoiem, D.: Where to look: Focus regions for visual question answering. In: CVPR (2016)
39. Song, S., Lan, C., Xing, J., Zeng, W., Liu, J.: An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In: AAAI (2017)
40. Thurau, C., Hlavác, V.: Pose primitive based human action recognition in videos or still images. In: CVPR (2008)
41. Wang, H., Kläser, A., Schmid, C., Liu, C.L.: Action recognition by dense trajectories. In: CVPR (2011)
42. Wang, H., Schmid, C.: Action recognition with improved trajectories. In: ICCV (2013)
43. Wang, Y., Jiang, H., Drew, M.S., Li, Z.N., Mori, G.: Unsupervised discovery of action classes. In: CVPR (2006)
44. Wörgötter, F., Aksoy, E.E., Krüger, N., Piater, J., Ude, A., Tamosiunaite, M.: A simple ontology of manipulation actions based on hand-object relations. IEEE Transactions on Autonomous Mental Development **5**(2), 117–134 (2013)
45. Xiao, T., Xu, Y., Yang, K., Zhang, J., Peng, Y., Zhang, Z.: The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In: CVPR (2015)

46. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A.C., Salakhutdinov, R., Zemel, R.S., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: ICML. vol. 14 (2015)
47. Yang, W., Wang, Y., Mori, G.: Recognizing human actions from still images with latent poses. In: CVPR (2010)
48. Yang, Y., Fermuller, C., Aloimonos, Y.: Detection of manipulation action consequences (mac). In: CVPR (2013)
49. Yang, Z., He, X., Gao, J., Deng, L., Smola, A.: Stacked attention networks for image question answering. In: CVPR (2016)
50. Yao, B., Fei-Fei, L.: Grouplet: A structured image representation for recognizing human and object interactions. In: CVPR (2010)
51. Yao, B., Fei-Fei, L.: Modeling mutual context of object and human pose in human-object interaction activities. In: CVPR (2010)
52. Yao, B., Fei-Fei, L.: Action recognition with exemplar based 2.5 d graph matching. In: ECCV (2012)
53. Yao, B., Jiang, X., Khosla, A., Lin, A.L., Guibas, L., Fei-Fei, L.: Human action recognition by learning bases of action attributes and parts. In: ICCV (2011)
54. Yao, B., Khosla, A., Fei-Fei, L.: Combining randomization and discrimination for fine-grained image categorization. In: CVPR (2011)
55. You, Q., Jin, H., Wang, Z., Fang, C., Luo, J.: Image captioning with semantic attention. In: CVPR (2016)