Distance-based Human-Object Interaction Detection

Weifeng Li1,*, Hongbing Yang1,a, Zhou Lei1,b, Dawei Niu2,c

1 School of Computer Engineering and Science, Shanghai University, Shanghai 200444, China
2 Zhejiang Uniview Technologies Co., Ltd
a email: hbyoungshu@staff.shu.edu.cn, b email: leiz@shu.edu.cn,
c email: niudawei@uniview.com

*Corresponding author’s e-mail: weify@shu.edu.cn

Abstract. In our lives, the interactions between humans and objects around them are happening all the time. The purpose of human-object interaction (HOI) detection is to locate humans and objects in a visual scene and infer the type of interaction between the two. Most of the HOI detection works infer the interaction type from two aspects: spatial features and visual features. We adopt the same strategy as them, but in the stage of extracting spatial features, we believe that the relationships between human body parts and objects are very important. We choose distance as a standard to measure the relationship between body parts and objects, and add it to the process of extracting spatial features, then use the refined spatial features to further extract visual features. Our experimental results on the V-COCO dataset show that our new model based on distance is effective. Compared with other methods, the accuracy of our model is significantly improved.

1. Introduction

The interactions between people and the surrounding environment are common. And with the development of multimedia equipment and technology, a large number of videos and pictures are produced every day. These pictures and videos contain numerous and different types of interactions. It is obviously unrealistic to rely on human handling for a large number of interaction instances. Fortunately, the development of deep learning provides the possibility of human-object interaction detection [1,2,3,4,6,7,8]. In recent years, as more and more scholars have devoted themselves to the field of deep learning, therefore deep learning has made considerable progress in computer vision fields such as object detection [5,9] and instance segmentation [10]. However, simply positioning and classifying the objects in images cannot effectively help the computer understand deeper information in images. We can regard the objects (including human and animals) in images as the basic elements that make up the scenes. Then the relationship between objects is one of the basic elements describing the content of images. In images, there are often various complicated relationships, and these relationships are the key to our understanding of images information. Human-Object interaction, as one of these relationships, has attracted more and more scholars’ attention in recent years. The purpose of Human-Object interaction detection is to detect human and objects in images and predicts the interaction types between them. By locating the position of the person-object pair, classifying the objects and the interactions, HOI detection can help the computers obtain more useful information.
from the images. Human-object interaction detection can be applied to many scenarios in life, such as scene monitoring, unmanned driving and intelligent robots.

The outputs of Human-Object Interaction detection can be expressed by \(<human, verb, object>\) triplets. "Human" and "object" represent instances of people and object, and "verb" represents predicted interaction label, such as “walk”, “look”. HOI detection \([2,4,7]\) can be divided into two stages. The first stage is to classify and locate the objects in the picture. Then, generating human-objects pairs bases on the human and object instances. In the second stage, the model will predict the interaction classes of each human-object pair. The prediction of interaction classes can be further divided into three aspects: human feature, object feature and spatial configuration. In the second stage, most HOI detection models \([1]\) can be seen as a three-branch structure: human branch, object branch and space branch. The human branch and the object branch are used to extract visual information related to interaction, and the spatial branch is responsible for extracting spatial information. The model will score the features of the three branches, merge the three different scores according to the score fusion strategy to obtain the final score, and complete interaction class inference. In Figure 1, we show two examples of human-object interaction.

2. Methods and Materials

There are many challenges in the field of HOI detection. For example, some interaction types (e.g. catch and hold) have only subtle differences. In addition, each person in the picture may interact with the objects around them. Therefore, even if there are only a few people and objects in the picture, there may be a lot of interaction. Most HOI detection work follows the following process. Given a picture, the object detector is used to detect all objects (including human and other objects) in the picture. Then the HOI detection model generates a set of human-object pair based on the detection results. For each pair, the model will extract features from three different streams, namely human stream, object stream and pairwise stream. The human feature stream extracts the human visual features from the picture according to the human bounding box, the feature will be used to predict the final HOI score. The same, object stream is used to extracts object feature. For third stream, pairwise stream used to extract spatial feature. There is no doubt about the importance of spatial features in predicting human-object interaction. Here is an example that can be used to illustrate the importance of spatial features. Imagine a picture where two people are riding bicycles, if the model only extracts the visual features of humans and bicycles to make predictions, we would get four different combinations of human and bicycle. Without considering the spatial feature, each combination has roughly the same probability of being predicted by the model as "riding a bike". This may cause the final detection accuracy to decrease. Therefore, spatial feature is very important in predicting the type of interaction. Recent years, many models obtain spatial feature map, which is used to further extract spatial features, based on human bounding box and object bounding box. In the spatial feature map, except for the places

![Figure 1](image-url)
surrounded by the bounding boxes of human and object, the values in other places are all zero. The final interaction prediction will be obtained through the fusion of the three streams.

2.1. Method
Unlike the work of other scholars, we believe that the distance between body parts and objects is the key basis for judging whether a human is performing a certain interaction. For example, when a person holds a cup, his hand must be close to the cup, and when a person eats an apple, his hand and mouth are closer to the apple. Therefore, we propose our model based on this idea. In order to obtain the distance between the body part and the object, we used the CPN [11] as a pose estimator to get the distribution of human joint points.

2.1.1. Object Detection and Human Pose Estimation
We choose Faster R-CNN[5] as object detector. Given an image, we use the object detector to obtain the human bounding boxes \( H = (b_1, b_2, ..., b_n) \) and object bounding boxes \( O = (b'_1, b'_2, ..., b'_m) \), and their score \( S_H = (s_1, s_2, ..., s_n) \), \( S_O = (s'_1, s'_2, ..., s'_m) \), where \( n \) denotes the total number of human instances, and \( m \) denotes the total number of object instances. In order to obtain the human keypoints, we use the CPN pose estimator [11] to generate the pose features \( H_K \) of the human body. \( H \) and \( O \) are the sets of bounding boxes of human and object respectively, and \( S_H, S_O \) are sets of their score. \( K_H \) is set of human keypoints, and for each \( k_h \in K_H, k = \{k^1_h, k^2_h, ..., k^N_h\} \), \( N = 17 \) is number of keypoints.

![Figure 2. This is our model structure. We use the combination of human pose and bounding boxes to extract spatial features in human-object interaction, and use it to extract the feature of human-object interaction pair. Finally, we get the final interaction score through multi-score fusion.](image)

2.1.2. Human-Object Pairwise Branch
Given an image, we use residual block to extract visual feature. In order to obtain human visual feature and object visual feature, we use convolution and pooling operations on according to the bounding box.

\[
f_h = (\text{Conv}(\text{RoI}(\Gamma, b_h)))
\]

\[
f_o = (\text{Conv}(\text{RoI}(\Gamma, b_o)))
\]
In order to extract the important information contained in the image, we further refine the context feature \( f_c \) using residual block.

\[
f_c = \text{Conv}(\Gamma)
\]  

(3)

For each human-object pair, we use \( b_a, b_o \) to generate the unified bounding box \( b_u \) of the two.

\[
b_u = b_a \cup b_o
\]  

(4)

In the same way, we can get \( f_u \).

\[
f_u = \text{Conv}(\text{RoI}(\Gamma, b_u))
\]  

(5)

According to the object position coordinates and the \( k_i \in K_H \) of human joint points, we can obtain the distance feature vector \( f_d \) through calculation. Because the distance between the body part and the object is the basis for judging the class of interaction, we merge \( f_u \) with the distance feature \( f_d \).

\[
f_u^f = f_u \otimes f_d
\]  

(6)

We fuse \( f_k, f_a, f_c, f_u \) to get human-object pairwise feature \( f_p \) and score \( s_p \).

\[
f_p = f_k \oplus f_a \oplus f_c \oplus f_u
\]  

(7)

Where \( \text{RoI} \) is the region of interest pooling, \( \text{Res} \) is the residual block, \( \text{GAP} \) is the global average pooling, \( \oplus \) is the concatenation operation, \( \otimes \) is element-wise multiplication.

2.1.3. Spatial Attention Branch

Most HOI detection work only uses the positions of human and objects when extracting spatial information, and these works ignore the importance of human pose in prediction. The spatial features extracted using only the bounding boxes of humans and objects are coarse. When we classify similar actions, the classification information provided by coarse spatial feature \( f_{bm} \) is very limited. Therefore, we add distance feature \( f_d \) to the process of extracting spatial features, and use the refined spatial features for pairwise features.

\[
f_s = f_{bm} \otimes f_d
\]  

(8)

\[
f_p = f_k \otimes f_p
\]  

(9)

We feed \( f'_j \) into two different network structures \( \text{Conv}_{ho} \) and \( \text{Conv}_p \), and their output sizes are 1 and 29 respectively. Finally, we can get two interaction scores \( s_{ho} \) and \( s_p \).

\[
s_{ho} = \sigma(\text{Conv}_{ho}(f'_j))
\]  

(10)

\[
s_p = \sigma(\text{Conv}_p(f'_j))
\]  

(11)

Where \( s_{ho} \) represents the possibility of interaction between human-object pair, and \( s_p \) represents 29 interaction scores between human-object pairs on V-COCO dataset.

2.1.4. Graph Convolutional Interaction Branch

Since there are multiple people or objects in an image, in order to better predict the interaction between them, we construct a graph. We take the human feature \( f_h \) and object feature \( f_o \) as the nodes of the graph, and the possibility of interaction \( s_{ho} \) between them as the adjacency value.

\[
v_{ho} = v_{oh} = s_{ho}
\]  

(12)

\[
f'_j = f_o + \sum_{j=1}^{k} v_{oh} M_{oh}(f_h)
\]  

(13)

\[
f_h = f_h + \sum_{i=1}^{k} v_{ho} M_{ho}(f_o)
\]  

(14)

Where \( v_{oh}, v_{ho} \) is adjacency values, and \( M_{oh}, M_{ho} \) is mapping function.

\[
f_p = \text{GCN}(f'_j \otimes f'_o)
\]  

(15)
2.1.5. Score Fusion
Therefore, we can get the interaction score.

\[ s_g = \sigma(Conv(f_g)) \]  

\[ s_s = \sigma(Conv(f_s)) \]  

Where \( s_s \) is spatial feature score and \( s_g \) is graph convolutional network feature score. Integrating all the scores, we can get the final interaction score \( s_f \).

\[ s_f = s_{hr} \cdot s_g \cdot s_s \cdot s_p \cdot s_p \cdot s_h \]  

2.2. Dataset and Metrics
This section is used to introduce experimental data, metrics.

2.2.1. Experimental data
We use V-COCO dataset to train, validate, and test the new model. V-COCO dataset. The V-COCO dataset is derived from the MS COCO dataset. The dataset consists of 10,346 images, of which 2,533 were used for training, 2,867 for validation, and 4,946 for testing. All the images in the dataset have a total of 16,199 person instances. And the dataset is annotated with 26 interaction categories. It should be noted that three of all interaction categories have two types of targets: “object” and “instrument”.

2.2.2. Evaluation metrics
We use the following method to evaluate the performance of our model. The result of the model is to detect human and object bounding boxes and the interaction labels between them, denoted as the \(< human, verb, object > \) triplet. The result is positive only when both detected human and object bounding boxes have \( IoUs \geq 0.5 \) with the ground-truth boxes, and detected interaction label matches the actual interaction type.

3. Result and Discussion
In order to verify the model, we train, verify and test the model on the V-COCO dataset. Compared with some previous work, our model has a great improvement in accuracy. In order to illustrate the important role of pose feature in extracting spatial features, in “Our-Pose” experiment, we used traditional method to extract spatial features, and the accuracy of the experimental results dropped by 0.64.

| Methods          | mAP role |
|------------------|----------|
| Gupta et al.[12] | 31.8     |
| InteractNet[6]   | 40.0     |
| GPNN[8]          | 44.0     |
| iCAN[4]          | 45.3     |
| Ours - Pose      | 46.27    |
| Ours             | 46.91    |

Table 2. Experimental results on V-COCO dataset

| HOI Class | mAP   |
|-----------|-------|
| hold-obj  | 46.27 |
| sit-instr | 26.85 |
| ride-instr| 67.05 |
| look-obj  | 42.81 |
| hit-instr | 76.24 |
| hit-obj   | 49.62 |
From our experimental results, it is not difficult to find the importance of pose feature in extracting spatial features. It's not difficult to understand. When human interacts with surrounding objects, the relative position between human and objects is important, but the information it can provide is relatively limited. And the relationship between human body parts and objects can provide more information for judging the type of interaction. In our model, we add the distance feature between body parts and objects into the spatial feature extraction process, and use it to extract visual features. This approach is undoubtedly a new innovation. The good news is that experiments prove that this method is feasible.

Acknowledgments
The authors would like to thank the Heigh Performance Computing Center of Shanghai University for providing the computing resources and technical support.

References
[1] Chao, Y., Liu, Y., Liu, X., Zeng, H. Deng, J. (2018) Learning to Detect Human-Object Interactions. In: 2018 IEEE Winter Conference on Applications of Computer Vision. Lake Tahoe, NV. pp.381-389.
[2] Ulutan, O., Iftekhar, A. S. M., Manjunath, B. S. (2020) VSGNet: Spatial Attention Network for Detecting Human Object Interactions Using Graph Convolutions. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, WA, USA. pp. 13614-13623.
[3] Wan, B., Zhou, D., Liu, Y., Li, R., He, X. (2019) Pose-Aware Multi-Level Feature Network for Human Object Interaction Detection. In: 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South). pp. 9468-9477.
[4] Gao, C., Zou, Y., Huang, J.-B. (2019) ICAN: Instance-centric attention network for human-object interaction detection. In: British Machine Vision Conference 2018.
[5] Ren, S., He, K., Girshick, R., Sun, J. (2017) Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137-1149.

[6] Gkioxari, G., Girshick, R., Dollár, P., He, K. (2018) Detecting and Recognizing Human-Object Interactions. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT. pp. 8359-8367.

[7] Gao, C., Xu, J., Zou, Y., Huang, J.B. (2020) DRG: Dual Relation Graph for Human-Object Interaction Detection. In: European Conference on Computer Vision. Glasgow, UK. pp.696-712.

[8] Qi, S., Wang, W., Jia B., Shen, J., Zhu, SC. (2018) Learning Human-Object Interactions by Graph Parsing Neural Networks. In: European Conference on Computer Vision. Munich, German. pp.407-423.

[9] Girshick, R., Donahue, J., Darrell, T., Malik, J. (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition. Columbus, United States. pp. 580–587.

[10] He, K., Gkioxari, G., Dollár, P., Girshick, R. (2017) Mask r-cnn. Proceedings of the IEEE international conference on computer vision. Venice, Italy. pp. 2961–2969.

[11] Chen, Y.; Wang, Z.; Peng, Y.; Zhang, Z.; Yu, G.; Sun, J. (2018) Cascaded pyramid network for multi-person pose estimation. Proceedings of the IEEE conference on computer vision and pattern recognition. Salt Lake City, UT, USA. pp. 7103–7112.

[12] S. Gupta and J. Malik, Visual semantic role labeling," arXiv preprint arXiv:1505.04474,2015.