Diagnosing Rarity in Human-object Interaction Detection

Mert Kilickaya
QUvA Deep Vision Lab
Amsterdam, Netherlands
m.kilickaya@uva.nl

Arnold Smeulders
QUvA Deep Vision Lab
Amsterdam, Netherlands
a.w.m.smeulders@uva.nl

Abstract

Human-object interaction (HOI) detection is a core task in computer vision. The goal is to localize all human-object pairs and recognize their interactions. An interaction defined by a <verb, noun> tuple leads to a long-tailed visual recognition challenge since many combinations are rarely represented. The performance of the proposed models is limited especially for the tail categories, but little has been done to understand the reason. To that end, in this paper, we propose to diagnose rarity in HOI detection. We propose a three-step strategy, namely Detection, Identification and Recognition where we carefully analyse the limiting factors by studying state-of-the-art models. Our findings indicate that detection and identification steps are altered by the interaction signals like occlusion and relative location, as a result limiting the recognition accuracy.

1. Introduction

The goal of HOI detection is to detect all possible human-object pairs and recognize their interactions from an image. It is a core task in computer vision with many applications in robotics [10]. A HOI is defined by a triplet of <human, interaction, object>, where the human and the object is a bounding box and the interaction is a <verb, noun> pair, such as <ride, bicycle>. The task received an increasing amount of attention in recent years [1, 2, 5, 7, 8, 13, 14, 17, 19] thanks to the benchmark dataset HICO-Det [3]. The distribution of the training samples for interactions follows a long-tailed distribution where many interactions have few examples, see Figure 1. Despite the progress in the performance of many-shot interactions, recognizing rare interactions remains a challenge.

HOI Detection is accomplished in three steps, see Figure 2. **1. Human-object detection**: HOI detector initially localizes all possible human-objects from the image. This step is challenged by the fact that interactions transform the human-object appearance, such as occlusions due to grasping, making it hard to localize all human-objects. **2. HOI Identification**: Given the exhaustive pairing of all detected humans and objects, the HOI detector needs to identify the real interacting pairs. In the case of Figure 2, only the rider, and the horse are in an interaction. This step is challenging since the cues of an interacting pair is in their subtlety, such as the relative locations of human-objects, human gaze or object parts. **3. HOI Recognition**: In this step, HOI detector needs to classify the interaction type of the human-object pair(s), such as <hold, cow> and <sit on, cow> in Figure 2. It is hard to distinguish among many interaction types since only a few examples are available for many interactions.

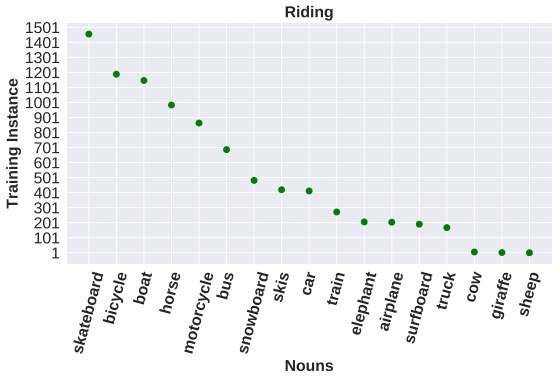

Figure 1: Human-object interactions exhibit a long-tail. For example, riding skateboard has more than $1k$ training instances whereas cow, giraffe or sheep only has 1 example.

Existing models highlight the difficulty in recognizing rare HOIs by reporting the performance on both the rare and nonrare (*i.e.* many shots) splits on the benchmark dataset [3]. However, it is not known what makes rare interactions particularly challenging aside from the low number of examples. Are the human-objects of rare interactions harder to detect? Or rare interacting pairs are harder to identify? The goal of our paper is to answer these questions. Instead of engineering a new model, we try to understand the detectability and identifiability of rare interactions with the

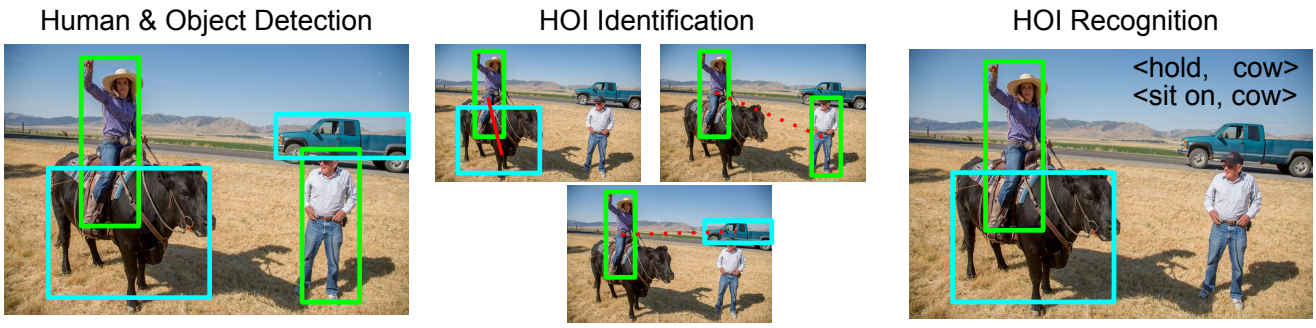

Figure 2: Steps of our diagnostic study. Detection (left), Identification (middle) and Recognition (right).

help of the state-of-the-art HOI detectors [3, 5, 11].

Our findings are: (i) Localizing human-objects of rare interactions is not challenging, however, detection is altered by the small and occluded human-objects, (ii) Identifying the rare HOIs is challenging and is altered by the background clutter and human-object distance, and (iii) Recognizing rare interactions is influenced by the detection and identification errors, leaving a big room for improvement.

2. Empirical Material

2.1. Benchmark Dataset

For our analysis, we resort to HICO-Det dataset [3]. HICO-Det is the biggest HOI detection benchmark, with a diverse set of categories. The dataset comes with (i) Human-object bounding boxes for detection, (ii) Human-object interaction pair annotations for identification, and (iii) Human-object interaction types for recognition. A typical human-object concurrently performs multiple interactions such as holding, sitting on and riding a bicycle which is exhaustively annotated. The dataset has in total $47k$ number of images, with more than $150k$ human-object pair annotations. There exists 600 distinct interaction types of which 168 are rare, for 80 unique nouns and 117 unique verbs.

A unique property of the dataset is, for each noun in the dataset, there exists a no-interaction category, where at least a human and the target object is in the image, though not performing an interaction(*e.g.* the man and the car in Figure 2). This enforces the models to focus on the interaction characteristics as opposed to leveraging human-object co-occurrence.

2.2. Benchmark HOI Detectors

For our diagnostic purposes, we resort to three state-of-the-art HOI detectors, namely HO-RCNN [3], iCAN [5], and TIN [11] for the following reasons: (i) Their high performance, (ii) Standard use of the same object detector, the same backbone and the same number of layers and (iii) Publicly available code. Here, we present an overview.

HO-RCNN [3]. HO-RCNN is a three-stream Convolutional Neural Network. Each stream considers the appearance of either the human, the object or the human-object (pairwise). Human and object streams consider the global appearance of human and object regions obtained via Region-of-Interest pooling [6], whereas pairwise stream considers the spatial layout of human-object locations. Spatial locations are critical especially to identify a possible interaction between a human and an object. The detector is built upon the Faster-RCNN backbone [15]. For human-object detection, the model makes use of this backbone pre-trained on MS-coco [12]. HOI recognition is achieved by combining the individual predictions of the three streams.

iCAN [5]. iCAN follows the same network structure as HO-RCNN, and couples HO-RCNN network with a self-attention mechanism [18] called instance-centric attention layer for the human and the object streams. This highlights the fine-grained details within human-object regions that are essential to HOIs.

TIN [11]. TIN augments iCAN with an interactivity classifier, that predicts whether if a given pair of human-object are in interaction or not. The authors prune (suppress) those pairs that are predicted to be non-interacting by the interactivity classifier prior to recognition. Since it is critical to identify which pairs are interacting before the recognition, the authors obtain a considerable improvement in both rare and nonrare interactions over iCAN.

Implementation details. All the models are trained for 1.8 million iterations with image-based training [15]. The learning rate is set to 0.001 decayed after $900k$ iterations. Weight decay (0.0001), dropout (with keep probability $p = 0.4$) and batch normalization [9] is used for regularization. We include all human detections with a confidence higher than 0.8 and all object detections with a confidence higher than 0.3. Using this standard gave a boost to HO-RCNN as it yields better results than iCAN in our experiments.

3. Diagnostic Analysis

We diagnose the HOI detection in three steps, namely Human-object detection, HOI identification, and HOI recognition. We use the standard rare/nonrare split proposed in [3].

3.1. Detection of Human and Object

To diagnose human and object detection, we first measure the recall of the off-the-shelf detector Faster-RCNN [15] commonly used by all three models. We measure the recall using the traditional PASCAL VOC criteria [4] that Intersection-over-Union IoU >0.50 over the full, the rare and the nonrare splits. Results can be seen from Table 1.

	Full	NonRare	Rare
Human	86	86	92
Object	63	64	90

Table 1: Recall of human and object detections.

Results show that there is no gap in terms of human and object detection performance between rare and nonrare categories. This indicates that recognizing rare interactions is not altered by the detection step.

Then, we study the sensitivity of the detector to the area and the occlusion. We measure the area of a bounding box by the width $\times$ height of the box divided by the number of pixels in the image. We measure the occlusion of box b_j on b_i as $\frac{b_i \cap b_i}{area(b_j)}$ [16]. An area or occlusion is small if <0.20. Results can be seen from Table 2.

	Full	Area		Occlusion	
		Small	Bigger	Small	Bigger
Human	86	44	94	95	66
Object	63	26	79	84	43

Table 2: Sensitivity of human-object detection.

Results indicate that the off-the-shelf detector is sensitive to both the small area and the large occlusion of humans and objects. This can limit the performance in subsequent steps since many human-object interactors occupy a small region in the image and is occluded by each other. We give detection examples in Figure 3 for the human and the objects (green for detected, red for missed boxes). Observe how, within the same image, the detector fails to localize the humans or the objects that have bigger occlusions or occupy a small region.

Figure 3: Example human-object detector success and failure cases.

It is concluded that localization of humans and objects is not altered for rare interactions, however, is affected by the area and the occlusion on human and object regions.

3.2. Identification of Human-Object Interaction

To diagnose HOI Identification, we measure the binary accuracy of identification (interacting vs. not-interacting) performance. To obtain interaction vs. not-interaction scores from each model, we obtain maximum response over all interaction (520 classes) and not-interaction (80 classes) categories respectively. Results are in Table 3.

	Full	NonRare	Rare
HO-RCNN [3]	79	79	71
iCAN [5]	74	75	67
TIN [11]	82	82	77

Table 3: Identification accuracy of human-object interaction for rare and non-rare interaction categories.

Results reveal that for all three models there is a consistent gap in the identification performance between rare and nonrare interactions. This indicates that the models demand more training examples to learn interactivity.

We then study the sensitivity of human-object distance and human-object clutter in identification accuracy. We compute human-object distance as the number of pixels between the centers of humans and objects divided by the number of pixels in the image. We compute human-object clutter as the number of human-objects in the background (*i.e.* not involved in any interaction). The distance is deemed to be small if <0.20 and the clutter deemed to be small if less than 5 other human-objects exist in the background. Results can be seen from Table 4.

	Full	Distance		Clutter	
		Small	Bigger	Small	Bigger
HO-RCNN [3]	79	94	77	91	69
iCAN [5]	74	91	72	88	65
TIN [11]	82	96	80	93	74

Table 4: Sensitivity of identification.

Results indicate that HOI identification is challenged by human-object distance. The models tend to assign distant human-objects to not-interaction category, leading to errors in identification. This shows that the models mostly leverage the spatial layout of human and object locations to identify the interaction.

It is also clear that background objects confuse the HOI identification step. As the number of possible human-object pairs increases with the background human-objects, the classifier finds it hard to distinguish interacting pairs from non-interacting ones.

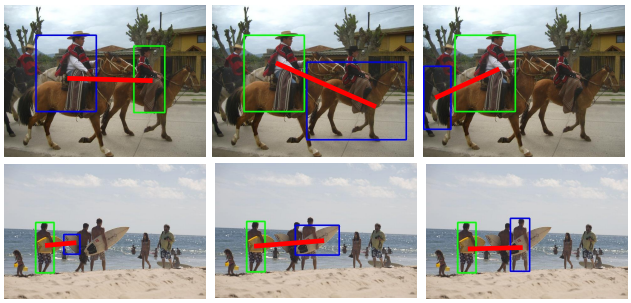

Figure 4: Failure examples for HOI Identification of TIN [11]. The rider (above) and the surfer (below) is paired with three other surrounding objects even though there is no interaction in between, lowering the identification performance.

The effect of the human-object distance and the human-object clutter on identification is visualized in Figure 4 for TIN [11]. In both images there are multiple humans and multiple objects, making it a real test for HOI identification. The rider (above) and the surfer (below) are paired with three other surrounding objects with even though there is no interaction in between, lowering the identification performance.

It is concluded that HOI identification is inferior for the rare interaction categories, and HOI identification is altered by the existence of the human-object distance and human-object clutter.

3.3. Recognition of Human-Object Interaction

To diagnose HOI recognition, in line with detection and identification steps, we measure the overall performance at the human-object instance-level. Specifically, we compute the mean Average Precision of interaction classification performance, which is then averaged over all pairs (not over classes) over the dataset. The results are presented in Table 5.

	Full	NonRare	Rare
HO-RCNN [3]	9	10	1
iCAN [5]	8	9	1
TIN [11]	10	10	2

Table 5: Recognition mAP for rare and non-rare interaction categories.

Results are in line with the class-based mAP reported in all three models [3, 5, 11], as there is a big gap in recognizing rare and nonrare interaction instances. We then measure the sensitivity of recognition to the detection and identification errors. A detection is correct if IoU >0.50, and incorrect otherwise. An identification is correct if the accuracy is 1. Results can be seen from Table 6.

	Full	Detection		Identification	
		Incorrect	Correct	Incorrect	Correct
HO-RCNN [3]	9	3	14	3	16
iCAN [5]	8	3	13	4	13
TIN [11]	10	4	14	4	23

Table 6: Sensitivity of recognition.

Results indicate that both the detection and identification errors alter the recognition performance as expected. However, a correct detection or identification does not guarantee perfect recognition accuracy, indicating that the human-object representation is not discriminative.

4. Conclusion

This paper focused on rarity in HOI detection in three steps. We revealed that human-object detection is altered by occlusions and area, that are abundant in HOIs. This calls for interaction-specific human-object detectors since for a conventional detector occlusion is a nuisance rather than a signal. We also showed that identifying rare HOIs is difficult which calls for finer-grained reasoning beyond spatial layout. Lastly, we show that recognition is limited by detection and identification errors, which leaves a big room for improvement for these specific steps.

References

[1] Ankan Bansal, Sai Saketh Rambhatla, Abhinav Shrivastava, and Rama Chellappa. Detecting human-object interactions via functional generalization. *arXiv preprint*, 2019. 1

[2] Sanaa Chafik, Astrid Orcesi, Romaric Audigier, and Bertrand Luvison. Classifying all interacting pairs in a single shot. *WACV*, 2020. 1

[3] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *WACV*, 2018. 1, 2, 3, 4

[4] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 2010. 3

[5] Chen Gao, Yuliang Zou, and Jia-Bin Huang. ican: Instance-centric attention network for human-object interaction detection. *BMVC*, 2018. 1, 2, 3, 4

[6] Ross Girshick. Fast r-cnn. In *ICCV*, 2015. 2

[7] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. Detecting and recognizing human-object interactions. In *CVPR*, 2018. 1

[8] Tanmay Gupta, Alexander Schwing, and Derek Hoiem. No-frills human-object interaction detection: Factorization, appearance and layout encodings, and training techniques. *arXiv preprint*, 2018. 1

[9] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015. 2

[10] Dov Katz and Oliver Brock. Manipulating articulated objects with interactive perception. In *ICRA*, 2008. 1

[11] Yong-Lu Li, Siyuan Zhou, Xijie Huang, Liang Xu, Ze Ma, Hao-Shu Fang, Yanfeng Wang, and Cewu Lu. Transferable interactiveness knowledge for human-object interaction detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3585–3594, 2019. 2, 3, 4

[12] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 2

[13] Julia Peyre, Ivan Laptev, Cordelia Schmid, and Josef Sivic. Detecting unseen visual relations using analogies. *arXiv preprint*, 2018. 1

[14] Siyuan Qi, Wenguan Wang, Baoxiong Jia, Jianbing Shen, and Song-Chun Zhu. Learning human-object interactions by graph parsing neural networks. In *ECCV*, 2018. 1

[15] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015. 2, 3

[16] Alexander Vezhnevets and Vittorio Ferrari. Object localization in imagenet by looking out of the window. *BMVC*, 2015. 3

[17] Tiancai Wang, Rao Muhammad Anwer, Muhammad Haris Khan, Fahad Shahbaz Khan, Yanwei Pang, Ling Shao, and Jorma Laaksonen. Deep contextual attention for human-object interaction detection. In *ICCV*, 2019. 1

[18] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, 2018. 2

[19] Bingjie Xu, Yongkang Wong, Junnan Li, Qi Zhao, and Mohan S Kankanhalli. Learning to detect human-object interactions with knowledge. In *CVPR*, 2019. 1