HICO: A Benchmark for Recognizing Human-Object Interactions in Images

Yu-Wei Chao, Zhan Wang, Yugeng He, Jiaxuan Wang, and Jia Deng
Computer Science and Engineering, University of Michigan, Ann Arbor
{ywchao,wangzhan,daranday,jiaxuan,jiadeng}@umich.edu

Abstract

We introduce a new benchmark "Humans Interacting with Common Objects" (HICO) for recognizing human-object interactions (HOI). We demonstrate the key features of HICO: a diverse set of interactions with common object categories, a list of well-defined, sense-based HOI categories, and an exhaustive labeling of co-occurring interactions with an object category in each image. We perform an in-depth analysis of representative current approaches and show that DNNs enjoy a significant edge. In addition, we show that semantic knowledge can significantly improve HOI recognition, especially for uncommon categories.

1. Introduction

Visual recognition of Human-Object Interactions (HOI) (e.g. "riding a bike", "boarding an airplane") in still images is a fundamental problem in computer vision. Successful HOI recognition is a prerequisite for generating rich image descriptions and retrieving images by sentences. HOI recognition is also an important subset of the larger problem of action and activity recognition, which also includes categories involving no direct interactions with objects, such as "walking", or high-level events, such as "concerts". [1]

HOI recognition differs from object/person recognition in that the key is to distinguish *a variety of different interactions with the same object category*. In other words, in addition to recognizing the presence of the a person and an object, it is critical to understand what the person is doing to the object. Is the person riding, walking, or repairing a bike? Is the person feeding, hunting, or watching a bear? Without an accurate understanding of the *interaction*, we will not be able to generate informative image descriptions besides a bag of objects.

Despite significant advances in recognizing humans [2] and objects [14], the state of the art of HOI recognition in images is still far from the demands of real-world applications. A key bottleneck is the limited number of HOI categories and limited interactions in current datasets.

In a literature survey, Guo and Lai [8] reported that the top-used dataset for still image based action recognition (including HOIs) between 2006 and 2013 is Pascal VOC 2010 [6], which contains only 9 categories. Stanford 40 Actions [35], which is the largest image-based action dataset before 2013, contains only 40 action categories. The recently released MPII Human Pose Dataset [1] contains annotated human poses from 410 human activities. While it is an excellent resource for human pose estimation and general action recognition, as will be analyzed in detail, it has limited diversity of interactions with each individual object category. Popular video-based action datasets, such as UCF 101 [28] and HMDB [15], share similar limitations. Without different interactions with the same object category, HOI recognition cannot be properly evaluated because a system can "cheat" by simply recognizing the objects.

Some image datasets are annotated with free-form texts [19, 17, 22, 37] that may include HOI descriptions. Although these datasets can in principle be used to evaluate HOI recognition, the utility is limited due to a number of fundamental challenges in computational linguistics. First, automatic extraction and parsing of phrases (e.g. verb-noun pairs) that describe human-object interactions are still unsolved. Second, even with the extracted verb-noun pairs, there is still the challenging problem of mapping words to meanings (senses). The same human-object interaction can be described in various forms (e.g. "repairs a bike", "doing bike repair", "fix a bicycle", "bike being repaired"), not to mention spelling errors. Thus while these datasets are ideal for evaluating image captioning, they are not suitable as benchmarks for HOI recognition—it would be hard to disentangle errors of language understanding and errors of HOI recognition.

In this paper, we introduce a new dataset for human-object interaction, "Humans Interacting with Common Objects" (HICO). It has a total of 47,774 images, covering 600 categories of human-object interactions (i.e. verb-object pairs such as "ride-bike") over 117 common actions (e.g. "ride", "feed", including one "no interaction" class) performed on 80 common objects (e.g.

[1] In this paper, we will use "actions" and "human-object interactions" interchangeably.

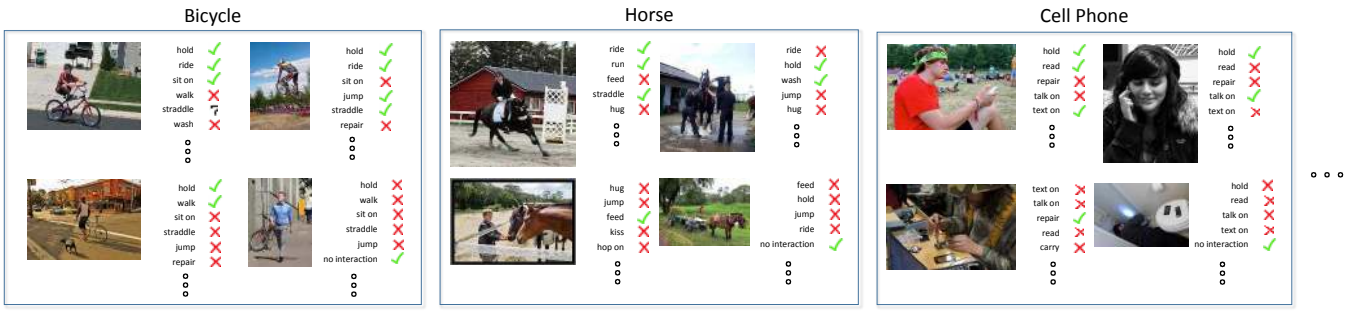


Figure 1: The "Humans Interacting with Common Objects" (HICO) dataset.

"bike", "bear"). The dataset is publicly available at http://www.umich.edu/~ywchao/hico/.

We highlight three key features of the dataset. First, it includes *diverse interactions* for each object category, i.e. an average of 6.5 distinct interactions per object category (not including the "no interaction" categories). Second, our HOI categories are *based on senses instead of words*. That is, we do not have "repair a bike" and "fix a bicycle" as separate categories, in contrast to the natural description based datasets such as [17]. Third, our annotations are *multi-labeled*, cognizant of the fact that different interactions with the same object often co-occur, e.g. "riding a bike and holding it" and "riding a bike but not holding it (hands-free)" are both plausible. Fig. 1 shows example images and annotations in HICO.

We demonstrate that our HICO dataset enables us to evaluate and analyze state of the art approaches on human-object interactions at a much larger scale. In particular, we study the following questions:

1. How well do the current state-of-the-art (action) classification approaches perform on HICO? Current approaches have only been tested on small datasets. It is unclear how they compare to each other on a dataset with a large number of action categories. Thus we compare a number of representative action recognition approaches including DNN-based methods.

2. Can semantic knowledge help recognizing uncommon human-object interactions? One challenge of HOI recognition at a large scale is that the data is highly unbalanced for different interactions. For example, "riding a bike" occurs much more frequently than "washing a bike". Here we investigate whether we can boost the recognition of uncommon classes by leveraging the semantic relations between the HOI classes (e.g. "wash dishes" and "wash a bike" share the action "wash") and co-occurrence knowledge.

The contributions of this work are two fold: (1) we introduce a new, publicly available dataset for recognizing human-object interactions, which enables targeted evaluation of HOI recognition at a large scale; (2) we perform in-depth analysis of current approaches, which sheds light on the challenges of large-scale HOI recognition and future research directions.

2. Constructing HICO

2.1. Selecting HOI Categories

The first step of constructing the dataset is to select a list of HOI categories. That is, we need a set of common objects and for each object, their respective common interactions. For common objects, we use the 80 object categories [2] introduced in the MS-COCO dataset [19], which were carefully selected based on children's vocabularies.

Next, we determine a set of common interactions for each object category. Since there is not an established list of "common" interactions, we take a language based approach by mining the actions described in the image captions of MS-COCO. Our assumption is that actions described in image captions will likely be more "visual" than those in a generic text corpus [20]. We use the Stanford Dependency Parser [27] to extract verbs appearing in the form of "verb-noun" or "verb-preposition-noun". To further expand the coverage of interactions, we also use the Google N-Gram dataset [20], which comes with dependency parsing results, and retrieve the top verbs that precede a particular object name. Since parsing is still a challenging NLP problem, we need to manually remove many incorrect results (e.g. "wear bike", "fly bike"). Combining the manually filtered results from MS-COCO and Google N-Gram, we obtain a set of candidate "common" verbs for each object category.

We then manually group the verbs with identical meanings into categories (e.g. "repair" is merged with "fix" in the context of "bicycle") and link them to nodes in WordNet (i.e. verb "synsets"). This is followed by a manual check that removes categories that are deemed too vague or abstract (e.g. "use", "take", "put"). The final selection is 520 verb-object pairs with 116 actions (verb senses) and 80 objects. For each of the 80 object categories, we add an extra "no interaction" category, e.g. "person is in the proximity but not interacting with bicycle". This gives a total of 600 HOI categories including the "no interactions" categories.

2.2. Image Collection and Annotation

To collect images for the HOI categories, we use Creative Common images from Flickr as the source of candidate images. Fig. 2 illustrates our annotation pipeline.

[2] A total of 91 categories exist but only 80 of them have annotations.

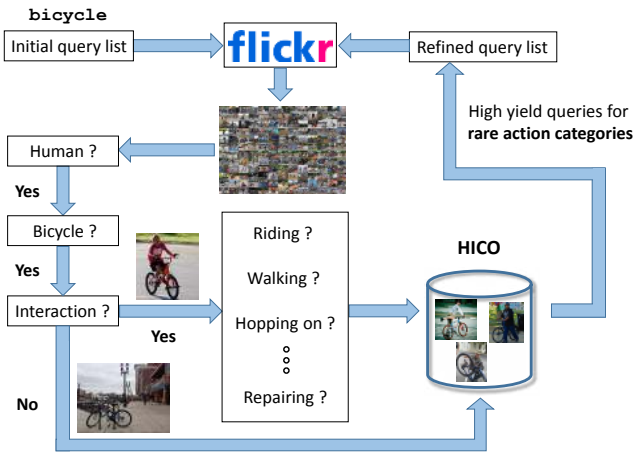


Figure 2: The pipeline of image collection and annotation.

We process each of the 80 object categories independently. Given an object category, e.g. "bicycle", we query Flickr with a barrage of related keywords ("bike", "fix bike", "fixing bike", "person bike" etc.).

Each retrieved candidate image from Flickr then goes through a series of annotation tasks on Amazon Mechanical Turk (AMT). For example, we first verify that the candidate image contains both "person" and "bike". If not, the image is discarded without further processing. Next we check whether there is an interaction at all. If none, then the image is marked as "person not interacting with bicycle". If yes, then each of the pre-defined interactions is checked individually. This then completes the annotation of one candidate image. To improve quality we have up to 3 workers answering each question and use a simple heuristic to combine answers. If there is any disagreement, the final answer is marked as "ambiguous/uncertain", which occurs a small fraction (9.77%) of the time and is often a result of ambiguity of interactions in images. For example, the annotation on "straddle bike" for the top-left image of Fig. 1 is an "ambiguous/uncertain" case.

For a given object category, we found that the distribution of images depicting different interactions is highly skewed. A few interactions dominate the candidate images (e.g. "ride", "hold", "sit on", "straddle" for "bicycle"), whereas the rest of the interactions (e.g. "walk", "hop on", "wash") are all very hard to come by. If left as is, this will create a challenge for benchmarking because too few images for the long tail categories will create too big a statistical variance for evaluation. To remedy this issue, we analyze the "yield" of each query keyword on the long tail interactions and use those high yield keywords to perform one more round of targeted image collection. Tab. 1 compares the statistics of candidate images in the second round of image collection to the initial round, showing a marked improvement of the percentage of uncommon interactions.

To summarize, each of the 80 common object categories is associated with a set of images. For each image in the set, it is guaranteed to contain both a human and the object category. It is also annotated exhaustively (in the form of "yes"/"no", occasionally "uncertain") for each of the possible human interations with this object category from a predefined list. Fig. 1 illustrates this structure [3]. It is worth noting that since each object is processed independently, we have performed deduplication and merged the duplicates between objects, i.e. some images (1.09%) are annotated with interactions with more than one object.

	#total	#no bike/person	#no inter	#ride	#repair
iter 1	2645	1369 (52%)	65 (2%)	763 (29%)	29 (1%)
iter 2	2561	1267 (49%)	149 (6%)	694 (27%)	51 (2%)

Table 1: Statistics of candidate images from Flickr in Iteration 1 and Iteration 2 (with targeted queries for rare interactions) for "bicycle". "no bike/person" means that the image has neither a person or a bike. "no inter" means that the image has a person and a bike but there is no interaction.

3. Related Datasets

We present a summary of related image datasets for action/HOI recognition in Tab. 2 [4]. Most existing work has been trained and evaluated on small-scale datasets such as PASCAL VOC Action Classification Challenge [6] and Stanford 40 Actions [35]. Our HICO dataset is one order of magnitute larger than these datasets in terms of both number of images and action categories. In the rest of this section we focus our discussion on several recent efforts towards scaling up action/HOI recognition.

TUHOI The "Trento Universal Human Object Interaction (TUHOI)" dataset [17] consists of 10,805 images over 2,974 actions. The images are from the ILSVRC 2013 [25] detection dataset and are annotated with action descriptions supplied by humans with no constraints on the vocabulary. On the surface our HICO dataset and TUHOI might look similar but there are two key differences.

First, our dataset has a restricted, sense-based category list (e.g. "fix a bike" and "repair a bike"' are the same category), whereas TUHOI is annotated with an open, word-based category list (e.g. "fix a bike" and "repair a bike" would count as separate categories). For this reason, 1,576 of the 2,974 TUHOI categories have only 1 image per category. Tab. 3 shows the interactions with the "bike" object in TUHOI, where "rid", "ridden", "ride", "ride a bike", "ride on", "rideon", and "riding" are counted as separate categories and most of them only have 1 image per category. In comparison, our "bike" interactions are grouped by senses and most have over 50 images (Tab. 4).

Second, for each object category, we exhaustively annotate all of its pre-defined interactions, i.e. we verify individ-

[3] In addition to our automatic pipeline, we also manually collected some images for categories with very few images.

[4] We omit video datasets since we focus on still image-based action recognition.

Dataset	#images	#actions	Sense	Clean
Sports event dataset [18]	1579	8	Y	Y
Ikizler *et al.* [11]	467	6	Y	Y
Ikizler-Cinbis *et al.* [12]	1727	5	Y	Y
The sports dataset [9]	300	6	Y	Y
Pascal VOC 2010 [6]	454	9	Y	Y
Pascal VOC 2011 [6]	2424	10	Y	Y
Pascal VOC 2012 [6]	4588	10	Y	Y
PPMI [33]	4800	12	Y	Y
Willow dataset [3]	968	7	Y	Y
Stanford 40 Actions [35]	9532	40	Y	Y
TBH dataset [23]	341	3	Y	Y
HICO (ours)	**47774**	**600**	**Y**	**Y**
89 action dataset [16]	2038	89	N	Y
TUHOI [17]	10805	2974	N	Y
MPII Human Pose [1]	40522	410	Y	Y
Google Image Search [24]	102830	2938	N	N

Table 2: Comparison of existing image datasets on action recognition. "Sense" means whether the category list is based on senses instead of words. "Clean" means whether the dataset is human verified.

ually "riding a bike", "holding a bike", and other interactions for the same image. Thus we are able to find images of "riding a bike but not holding it" and to pull out accurate co-occurrence statistics of the interactions. In contrast, TUHOI does not have this exhaustive verification. That is, the absence of "holding a bike" does not mean that the person is in fact not holding the bike—it could simply be that the annotator did not bother to mention it. Thus the annotations of TUHOI are affected by what annotators *choose* to describe.

For the above reasons, our dataset and TUHOI are *complementary* to each other. The annotations in TUHOI are ideal for benchmarking the task of generating *natural action descriptions* as would be provided by humans. Our dataset, on the other hand, insulates HOI recognition from the nuances of language expressions and the complex process of humans choosing what is worth describing.

MS-COCO MS-COCO [19] has a large number of high quality annotations of object segmentation masks and image captions. Although MS-COCO is not designed for action recognition, the verb phrases in the captions can in principle be extracted to evaluate actions. In fact, we have done so semi-automatically in order to determine the list of our HOI categories. However, one issue is that the current form of MS-COCO does not have enough images for many long tail categories. Fig. 3 compares the distribution of "bicycle" interactions in our dataset versus those extracted from MS-COCO captions. This shows that without targeted collection of new images for the long tail categories, MS-COCO in its current form is less suitable for HOI benchmarking. Another issue is that as with TUHOI, MS-COCO captions have the same complication of human bias in terms of what is worth describing.

verb tag	#	verb tag	#	verb tag	#	verb tag	#
bike	1	move	3	rid	2	sit on	10
cycle	7	na	1	ridden	1	stand	9
flip	1	no	1	ride	272	stand beside	1
freewheel	2	park	1	ride a bike	1	stand by	1
guide	1	pedal	5	ride on	6	stand near	1
hang	2	peddle	5	ridenon	1	steer	1
hold	21	play	3	riding	1	stop	1
hold up	1	prop	1	roll	1	touch	3
jump	2	push	5	sat	1	tricks	1
lean	1	race	4	sit	10	walk	1
look	1	repair	1	sit by	1	walk next to	1
						walk with	1

Table 3: Interactions with bicycle in the TUHOI dataset.

verb	#im	definition
carry, transport	30	move while supporting, either in a vehicle or in one's hands or on one's body
hold, take hold	1392	held by hand; to have or maintain in the grasp; to attach the hand to
inspect	124	to look at (something) carefully in order to learn more about it, to find problems, etc.
jump, leap	150	cause to jump or leap
hop on, mount, mount up, get on, jump on, climb on, bestride	26	climb up onto; get up on the back of
park	18	place temporarily
push, force	117	move with force, "He pushed the table into a corner"
repair, mend, fix, bushel, doctor, furbish up, restore, touch on	89	restore by replacing a part or putting together what is torn or broken
ride	1460	sit on and control a vehicle
sit on	1197	be seated
straddle	1511	sit or stand astride of
walk	187	to accompany on foot; to cause to move by walking
wash, rinse	6	clean with some chemical process
no interaction	174	

Table 4: Interactions with bicycles in our dataset.

MPII Human Pose The MPII Human Pose Dataset [1] (MPII in short) is a large-scale benchmark for 2D human pose estimation and action recognition. It has 40,522 images and 410 action categories. While similar in scale, its selection of categories is geared toward covering common daily activities instead of delineating different interactions with each object category. Tab. 5 compares the number of HOI categories (those taking a verb-noun form in its definition) of MPII with our dataset: MPII has on average 1.55 different interactions per object category whereas HICO has 6.5 (not including the "no interaction" categories). Thus our dataset is more suitable for evaluating diverse interactions with the same object categories.

Google Image Search Ramanathan et al. [24] reported results on 27K action categories. However, only a subset of

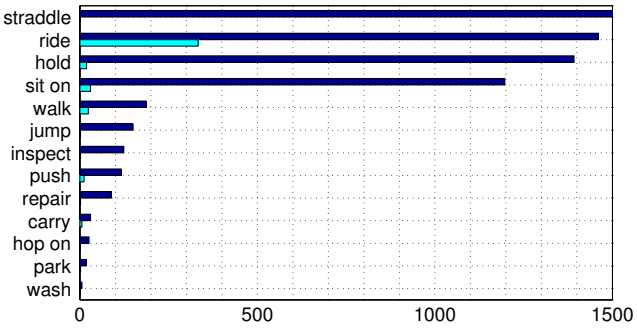


Figure 3: Number of images for each HOI category for "bicycle" (cyan: MS COCO, blue: our HICO dataset).

the data, 2,938 actions and 102,830 images, are made publicly available. For each action category, this subset contains the top 35 images returned by Google Image Search, which are treated as ground truth positives without any human verification. The lack of manual clean-up makes it less authoritative as our dataset for fine-grained comparisons of different approaches.

4. Benchmarking Representative Approaches

In this section we evaluate a few representative approaches for action recognition on our HICO dataset. We start with a brief review of prior work.

4.1. Related Work

Recognizing actions or human-object interactions in still images has been an actively studied topic. Prior work has explored a variety of methods and strategies. Some work exploits the special role of human poses (i.e. detecting human body parts) by deriving feature representations based on human poses [11, 38, 21, 30, 32, 31], whereas others focus more on spatial relations between humans and objects [34, 33, 23, 9, 5, 4]. Another line of work recognizes the fine-grained nature of action recognition and leverages discriminative templates [36, 26], color [13], or exemplars [10]. Recent work by Ramanathan et al. [24] showed action recognition can also be improved by exploiting the semantic relations.

4.2. Evaluation Setup

We use mean average precision (mAP) as our evaluation metric. Given an image, an approach being evaluated outputs a classification score for each of the HOI categories. Then we compute the average precision (AP) for each HOI category by ranking the test images by the classification scores. The average of AP for all HOI categories gives the mAP. This evaluation metric is motivated by the fact that many HOI categories are not mutually exclusive. This is similar to the metric used by PASCAL VOC Classification Competition [6], where the object classes can co-occur in images.

	#action	#HOI	#object	#action/object
MPII Human Pose [1]	410	102	66	1.55
HICO (ours)	**520**	**520**	**80**	**6.50**

Table 5: Comparison of action/HOI categories between MPII Human Pose [1] and our dataset (excluding "no interaction" classes).

In computing the AP for each HOI category, there is a subtlety in what test images we treat as ground truth negatives. Recall that for each HOI category (let's use "riding a bike" as a running example), the test images can be put into four groups:

- Verified positives: those verified to be "riding a bike";
- Verified negatives: those verified to contain a person and a bike but no "person riding bike";
- Ambiguous/Uncertain images: those verified to contain a person and a bike, but with disagreements among crowd workers on whether there is "person riding bike";
- "Unknown" images: those verified to contain a person and *some other object category*, e.g. "cat".

One setting is to use the verified positives as positives, skip the ambiguous image and the "Unknown" images, and use the verified negatives as negatives. This is equivalent to assuming that we will be able to perfectly filter out images with no "bicycles" before trying to recognize the interactions. We refer to the easier setting as the "Known Object (KO)" setting.

The "Known Object" setting might be "too easy" and unrealistic in the sense that a recognition system will not have a chance to be distracted by images *not* containing the correct object category. We thus add a more realistic setting by treating the "Unknown" images as extra ground truth negatives. This is closer to a realistic setting where there is no prior knowledge on what objects are present in a test image. Although there is the chance that some "Unknown" images may actually contain "person riding bike" (thus corrupting the evaluation), we have checked that the risk is small enough to be acceptable: we randomly sampled 10 HOI categories, manually went through all of their "Unknown" images, and found only 0.3% of them to be positives. In this paper, unless otherwise noted, all evaluations default to this more realistic setting.

We use a 80-20 training-test split, with the additional constraint that each HOI category should have at least 5 test images. In our dataset all HOI categories have at least 6 images, so all categories have at least 1 training images. This creates a one-shot learning setup for a subset of categories (51 out of 600), which is a natural result of the long tail distribution of HOI categories and is a challenge any practical HOI recognition approach must address.

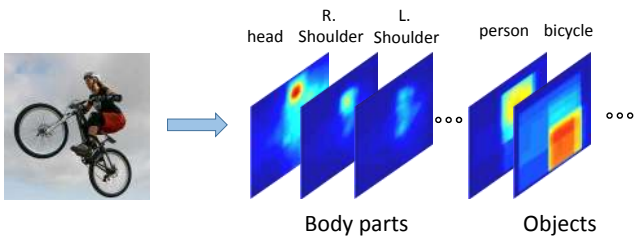


Figure 4: Heatmaps of object detection and human pose estimation as input to HOCNN.

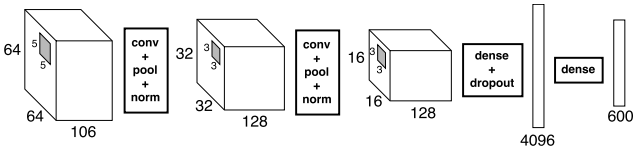


Figure 5: Network architecture of HOCNN.

4.3. Representative Approaches

We benchmark the following approaches on our new HICO dataset.

RandomForest (RF) [36] This approach won both the PASCAL VOC 2011 and PASCAL VOC 2012 Action Classification Competitions [6]. It uses random forests to select discriminative image regions, represented using SIFT features. We include it here to represent the state of the art of action/HOI recognition in static images.

FisherVectors (FV) [29] Prior to the breakthrough achieved by ConvNets [14], Fisher Vectors based approaches were the state of the art on image classification and won the ILSVRC2011 competition [25]. It is selected to represent traditional generic image classification approaches. A binary SVM is trained for each HOI category.

DNN We use DNN features from Alex's Net [14] pretrained on ImageNet and learn one binary SVM per HOI category. This represents the current state of the art approach in image classification. We also evaluate the following variants with different ways of fine-tuning.

- *Fine-tune V*. Fine-tune Alex's Net to classify only verb categories (i.e. group "wash bike", "wash car", "wash cat", etc. all into one "wash" category). This is to learn features that are common to a particular action such as "wash" regardless the objects.
- *Fine-tune O*. Fine-tune Alex's Net to classify only object categories (i.e. group "wash bike", "repair bike", "ride bike", etc. all into one "bike' category). This is to learn features that are common to a particular object.
- *Fine-tune VO*. Fine-tune Alex's Net to classify the verb-object pairs, i.e. directly the HOI categories.

	mAP	mAP (KO)
Random	0.57	33.37
RF [36]	7.30	38.15
FV [29]	4.21	37.74
DNN (ImageNet)	18.58	48.22
DNN (fine-tune V)	17.65	**49.07**
DNN (fine-tune O)	**19.38**	47.42
DNN (fine-tune VO)	18.08	47.89
HOCNN	4.90	39.05

Table 6: Performances of representative approaches.

HOCNN (Human-Object CNN) In this approach, we use the outputs of object detection and human pose estimation as features, on top of which we learn a Convolutional Neural Network (CNN) to classify the HOI categories. Specifically, given an input image, we first run object detectors and a pose estimator, which together generate a set of heatmaps, one per object category and one per human body part. A total of 106 heatmaps (80 object categories plus 26 body parts) are stacked together as the input to a CNN architecture (Fig. 4 and Fig. 5).

This approach serves to shed light on the role of the spatial relations between humans and objects for HOI recognition. It has been common for prior work to design explicit feature representations based on human-object spatial relations [33, 23, 9, 5, 4]. The question we ask here is to what extent such a mid-level representation can help HOI recognition, especially with a more powerful learning tool such as CNN.

To generate the input heatmaps, we train R-CNN [7] object detectors for the 80 object categories using the images in MS-COCO. [5] and use the pre-trained human pose estimator developed by Chen et al. [2].

Results Tab. 6 presents the mAP of the aforementioned approaches in both the default and the "Known Object" settings. In the default setting, DNN based approaches overwhelmingly outperformed traditional approaches (RandomForest and FisherVectors). Although the ordering is not surprising given the recent success of DNNs, the large gap (18.58 mAP versus 7.30 and 4.21 mAP) is still somewhat shocking. Unsurprisingly, RandomForest (RF) outperforms FisherVectors (FV), as the former was specifically designed for action recognition. Among the fine-tuned DNN variants, fine-tuning for objects (fine-tune O) achieves the best mAP in the default evaluation setting, suggesting that the more prevalent source of error is on recognizing the objects.

In the "Known Object" setting, however, fine-tuning for verbs (V), not objects, gives the best result. This is consistent with the fact that objects are no longer a source of error in this setting. It is also worth noting that, in this setting, all methods are not that much better than random chance, suggesting that the key challenge of HOI recognition – *rec-*

[5] We found 1776 images (out of 45,786) in HICO that are also in MS-COCO. These duplicate images are put into the training split of HICO to ensure test images of HICO are not used in any training.

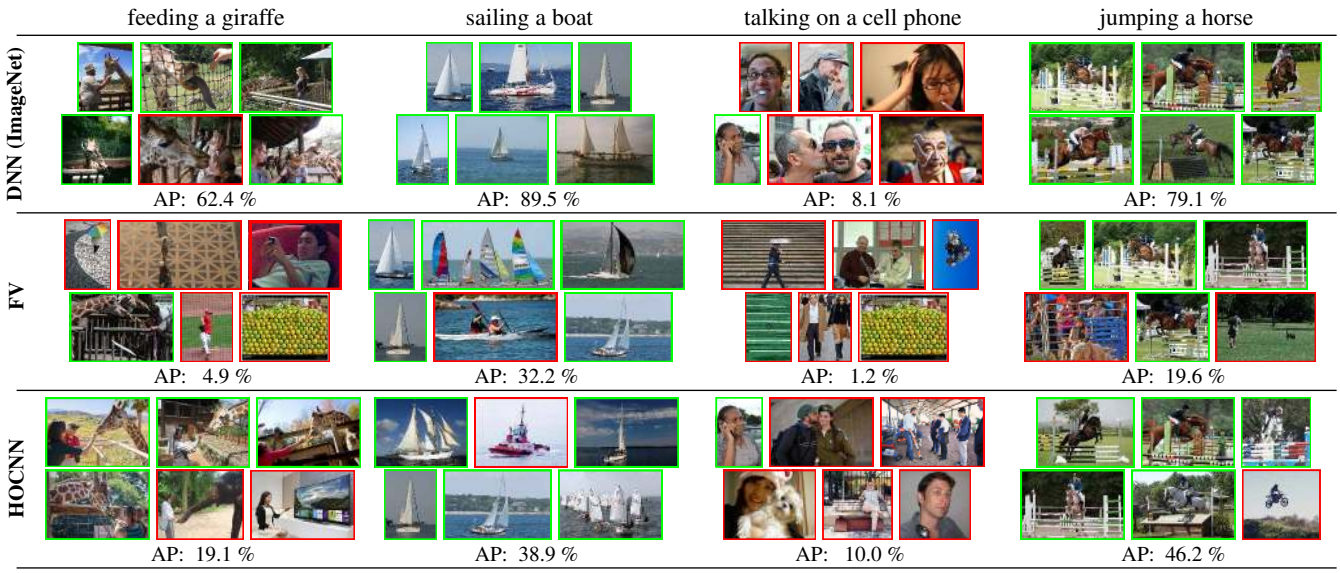


Figure 6: Top ranked images for different HOI categories in the default setting. Each row (column) represents one representative approach (HOI category). Green and red boxes represent ground-truth positives and negatives, respectively.

ognizing the interaction is still largely unsolved, even with DNNs.

HOCNN, although not performing as well as RandomForest (RF) in the default setting, outperforms both RandomForest (RF) and FisherVector (FV) in the "Known Object" setting. This suggests that human-object spatial relations are important in recognizing different interactions, and the high mAP of RF in the default setting should be attributed to better object recognition. We also observe that in both settings, there is still a large gap between HOCNN and DNNs that take pixels as input. This can possibly be attributed to the fact that end-to-end DNNs have the flexibility to select and combine all types of cues, from low level to high level and from local to global, whereas HOCNN is restricted to spatial relations between humans and objects.

Fig. 6 shows the top ranked images returned by DNN, FV, and HOCNN for a few HOI categories in the default setting. These examples suggest that all the approaches perform better for HOI categories with salient objects (e.g."sailing a boat", "jumping a horse"). AP drops significantly when the objects are smaller and harder to detect (e.g. "talking on a cell phone"). We also observe that, even when the objects can be reliably detected, distinguishing the interactions is still very challenging. For example, DNN (ImageNet), the best performing approach, cannot tell "kissing a giraffe" from "feeding a giraffe".

4.4. Using Semantic Knowledge

Knowledge on Compositions As discussed earlier, a major challenge of HOI recognition is the categories in the long tail that have very few training images. Humans have no difficulty recognizing these categories: even if we have not seen "washing a bike", we can still recognize the interaction with ease. This is because we likely have seen

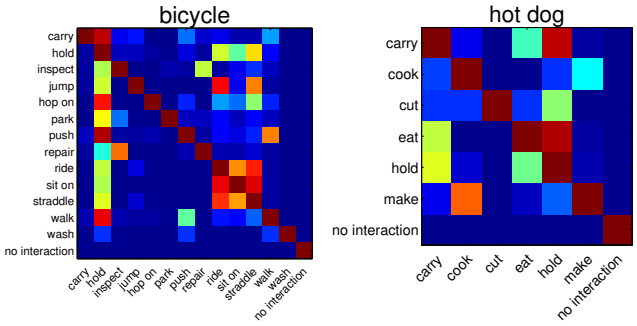


Figure 7: Cooccurrences of interactions.

common interactions such as "washing dishes" and have understood the concept of "washing" independent of the object being washed. In our HICO dataset, there is significant sharing of actions/verb senses between HOI categories: on average, each action/verb sense is paired with 4.5 different object categories. HICO thus provides an opportunity to test the hypothesis that semantic knowledge can be used to improve the recognition of rare HOI categories.

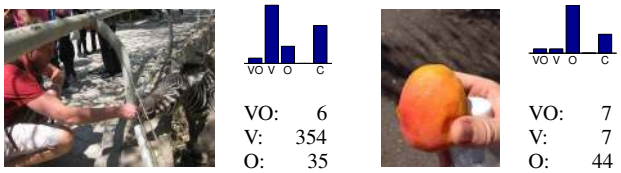
We experiment with the simplest possible strategy. Using the training data of HICO, we learn three types of classifiers, those for classifying verb-object (VO) pairs, those for verbs (V) only, and those for objects (O) only. This is similar to the fine-tuning of DNNs in Sec. 4.3. Then we explore various combinations of the three types of classifiers. For example, if a particular VO pair (e.g. "feeding a zebra") has very few training images (and as a result a weak VO classifier), we can instead learn a new classifier for "feeding a zebra" by combining the outputs of a V classifier trained to recognize"feeding" regardless of objects and an O classifier trained to recognize "zebra" regardless of the verbs.

Knowledge on Cooccurrences We also evaluate whether another type of knowledge, namely the cooccurrences of

	VO		V+O		V+VO		O+VO		V+O+VO		VO+coocc		V+O+VO+coocc	
	F	R	F	R	F	R	F	R	F	R	F	R	F	R
Random	0.57	0.18	0.57	0.18	0.57	0.18	0.57	0.18	0.57	0.18	0.57	0.18	0.57	0.18
DNN (ImageNet)	18.58	**0.74**	18.42	5.04	18.64	**1.73**	20.95	**5.07**	20.71	4.98	20.13	4.18	21.06	5.72
DNN (fine-tune V)	17.65	0.29	17.47	4.75	16.94	1.42	19.41	4.67	18.86	3.96	18.76	3.64	19.26	5.45
DNN (fine-tune O)	**19.38**	0.39	**18.68**	**5.99**	**19.43**	1.31	**21.52**	4.70	**21.33**	**5.07**	**20.91**	**4.31**	**21.66**	**5.91**
DNN (fine-tune VO)	18.08	0.39	17.44	4.79	18.41	1.68	19.36	4.40	19.38	4.51	18.98	3.94	19.62	5.35
HOCNN	4.90	0.16	5.40	0.51	5.09	0.21	5.38	0.32	5.47	0.32	5.18	0.32	5.51	0.41

	VO		V+O		V+VO		O+VO		V+O+VO		VO+coocc		V+O+VO+coocc	
	F	R	F	R	F	R	F	R	F	R	F	R	F	R
Random	33.37	19.26	33.37	19.26	33.37	19.26	33.37	**19.26**	33.37	19.26	33.37	19.26	33.37	19.26
DNN (ImageNet)	48.22	**23.41**	49.40	30.76	51.52	32.66	46.59	16.94	49.15	20.61	47.97	22.32	49.11	20.04
DNN (fine-tune V)	**49.07**	22.80	**49.98**	31.81	**51.56**	**33.86**	**48.04**	17.58	**49.58**	**25.52**	**49.08**	**22.85**	**49.31**	**23.80**
DNN (fine-tune O)	47.42	22.12	47.97	27.37	50.68	32.00	45.83	16.58	47.87	19.23	47.14	22.08	47.65	18.23
DNN (fine-tune VO)	47.89	22.17	49.87	**32.46**	50.51	32.78	46.81	17.02	48.30	20.99	47.76	21.79	48.17	20.30
HOCNN	39.05	20.81	39.74	21.79	40.74	23.09	38.15	18.12	39.80	19.51	38.81	20.29	39.72	19.57

Table 7: Performance of different combinations of V, VO, and O classifiers with multiple feature representations. Top: default setting, Bottom: "Known Object" setting. Performance measured as mAP on all 600 HOI classes (F) and 167 rare classes (R)—those with less than 5 positive training examples.



(a) feeding a zebra (b) squeezing an orange

Figure 8: Top-right: prediction scores of the verb-object (VO) pair, verb (V), object (O), and the combination (C). Bottom-right: number of training samples for VO/V/O.

actions, can improve HOI recognition. The intuition is that a rare category might piggyback on an cooccurring interaction that has more training data and easier to recognize. Our HICO dataset provides an ideal setting to test this hypothesis because of co-occurring interactions are exhaustively annotated. For example, Fig. 7 shows the co-occurrences of interactions for a few categories in HICO.

Again we evaluate the simplest possible method: we learn a new classifier to combine the outputs of all VO classifiers might co-occur on the same object category. For example, suppose we have trained the (binary) VO classifiers for "eating a hot dog", "holding a hot dog", and "riding a bike". Then we learn a new VO classifier for "eating a hot dog" that combines the outputs of the original two VO classifiers for "eating a hot dog" and "holding a hot dog"—"riding a bike" is not used because it is not a VO classifier on the same object.

It is worth noting that in all of our experiments no test images are ever used to learn any new classifiers that combine the outputs of existing classifiers. All learning is done using only the training set and cross-validation is used to prevent overfitting. Please refer to the supplemental material for our detailed setup.

Results Tab. 7 (top) summarizes the results in the default test setting (with extra negatives). It shows that regardless of feature representations, adding the V classifiers leads to moderate but consistent improvement for overall mAP as well as mAP for rare classes (e.g.for DNN (ImageNet), from 18.58 to 18.64 on the full dataset and from 0.74 to 1.73 on the rare action categories). The biggest improvements come from adding the O classifier, especially on the rare classes. In addition, adding cooccurrence knowledge consistently improves performance. Fig. 8 presents the predictions of the different types of classifiers on example images, illustrating how V classifiers and O classifiers help when a VO classifier is trained with very few images. The best result is achieved by combining the compositional knowledge and the co-occurrence knowledge (V+O+VO+coocc).

Tab. 7 (bottom) presents an evaluation of the same algorithms in the "Known Object (KO)" setting, as described in Sec. 4.2. This setting assumes zero errors of object recognition and focuses the evaluation on recognizing the interactions. We see that adding O classifiers to any combintation significantly hurts performance, which is expected given that the objects are already recognized and adding O classifiers will only cause overfitting. Another notable change in this setting is that adding V classifiers (compositional knowledge) leads to much more pronounced improvements (e.g. an increase of 9.25% absolute in mAP from VO to V+VO for DNN(ImageNet) on rare categories). This underscores the promise of leveraging semantic knowledge for large-scale HOI recognition.

5. Conclusions

We have introduced a new benchmark "Humans Interacting with Common Objects" (HICO) for recognizing human-object interactions (HOI). We have demonstrated the key features of our dataset: a diverse set of interactions with common object categories, a list of well-defined, sense-based HOI categories, and an exhaustive labeling of co-occurring interactions with an object category in each image. We have performed an in-depth analysis of representative current approaches and shown that DNNs enjoy a significant edge. In addition, we have shown that semantic knowledge can significantly improve HOI recognition, especially for uncommon categories.

Acknowledgement This work is partially supported by research awards from Google and Yahoo, and a hardware donation from Nvidia.

References

[1] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR*, 2014.

[2] X. Chen and A. L. Yuille. Articulated pose estimation by a graphical model with image dependent pairwise relations. In *NIPS*. 2014.

[3] V. Delaitre, I. Laptev, and J. Sivic. Recognizing human actions in still images: a study of bag-of-features and part-based representations. In *BMVC*, 2010.

[4] V. Delaitre, J. Sivic, and I. Laptev. Learning person-object interactions for action recognition in still images. In *NIPS*. 2011.

[5] C. Desai, D. Ramanan, and C. Fowlkes. Discriminative models for static human-object interactions. In *CVPR Workshop on Structured models in Computer Vision*, 2010.

[6] M. Everingham, S. Eslami, L. Van Gool, C. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, 2015.

[7] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.

[8] G. Guo and A. Lai. A survey on still image based human action recognition. *Pattern Recognition*, 47(10):3343 – 3361, 2014.

[9] A. Gupta, A. Kembhavi, and L. Davis. Observing human-object interactions: Using spatial and functional compatibility for recognition. *TPAMI*, 31(10):1775–1789, Oct 2009.

[10] J.-F. Hu, W.-S. Zheng, J. Lai, S. Gong, and T. Xiang. Recognising human-object interaction via exemplar based modelling. In *CVPR*, 2013.

[11] N. Ikizler, R. Cinbis, S. Pehlivan, and P. Duygulu. Recognizing actions from still images. In *ICPR*, 2008.

[12] N. Ikizler-Cinbis, R. Cinbis, and S. Sclaroff. Learning actions from the web. In *ICCV*, 2009.

[13] F. Khan, R. Muhammad Anwer, J. van de Weijer, A. Bagdanov, A. Lopez, and M. Felsberg. Coloring action recognition in still images. *IJCV*, 105(3):205–221, 2013.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*. 2012.

[15] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. Hmdb: A large video database for human motion recognition. In *ICCV*, 2011.

[16] D. T. Le, R. Bernardi, and J. Uijlings. Exploiting language models to recognize unseen actions. In *ICMR*, 2013.

[17] D.-T. Le, J. Uijlings, and R. Bernardi. Tuhoi: Trento universal human object interaction dataset. In *Proceedings of the Third Workshop on Vision and Language*, 2014.

[18] L.-J. Li and L. Fei-Fei. What, where and who? classifying events by scene and object recognition. In *CVPR*, 2007.

[19] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollr, and C. Zitnick. Microsoft coco: Common objects in context. In *ECCV*. 2014.

[20] Y. Lin, J.-B. Michel, E. L. Aiden, J. Orwant, W. Brockman, and S. Petrov. Syntactic annotations for the google books ngram corpus. In *Proc. ACL 2012 System Demonstrations*, 2012.

[21] S. Maji, L. Bourdev, and J. Malik. Action recognition from a distributed representation of pose and appearance. In *CVPR*, 2011.

[22] V. Ordonez, G. Kulkarni, and T. L. Berg. Im2text: Describing images using 1 million captioned photographs. In *NIPS*, pages 1143–1151, 2011.

[23] A. Prest, C. Schmid, and V. Ferrari. Weakly supervised learning of interactions between humans and objects. *TPAMI*, 34(3):601–614, March 2012.

[24] V. Ramanathan, C. Li, J. Deng, W. Han, Z. Li, K. Gu, Y. Song, S. Bengio, C. Rossenberg, and L. Fei-Fei. Learning semantic relationships for better action retrieval in images. In *CVPR*, Boston, MA, USA, June 2015.

[25] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. Imagenet large scale visual recognition challenge, 2014.

[26] G. Sharma, F. Jurie, and C. Schmid. Expanded parts model for human attribute and action recognition in still images. In *CVPR*, 2013.

[27] R. Socher, J. Bauer, C. D. Manning, and N. Andrew Y. Parsing with compositional vector grammars. In *ACL*, 2013.

[28] K. Soomro, A. R. Zamir, and M. Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, 2012.

[29] J. Snchez, F. Perronnin, T. Mensink, and J. Verbeek. Image classification with the fisher vector: Theory and practice. *IJCV*, 105(3):222–245, 2013.

[30] C. Thurau and V. Hlavac. Pose primitive based human action recognition in videos or still images. In *CVPR*, 2008.

[31] Y. Wang, H. Jiang, M. Drew, Z.-N. Li, and G. Mori. Unsupervised discovery of action classes. In *CVPR*, 2006.

[32] W. Yang, Y. Wang, and G. Mori. Recognizing human actions from still images with latent poses. In *CVPR*, 2010.

[33] B. Yao and L. Fei-Fei. Grouplet: A structured image representation for recognizing human and object interactions. In *CVPR*, 2010.

[34] B. Yao and L. Fei-Fei. Modeling mutual context of object and human pose in human-object interaction activities. In *CVPR*, 2010.

[35] B. Yao, X. Jiang, A. Khosla, A. Lin, L. Guibas, and L. Fei-Fei. Human action recognition by learning bases of action attributes and parts. In *ICCV*, 2011.

[36] B. Yao, A. Khosla, and L. Fei-Fei. Combining randomization and discrimination for fine-grained image categorization. In *CVPR*, 2011.

[37] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *ACL*, 2:67–78, 2014.

[38] Y. Zheng, Y.-J. Zhang, X. Li, and B.-D. Liu. Action recognition in still images using a combination of human pose and context information. In *ICIP*, 2012.