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

Yu-Wei Chao, Zhan Wang, Yugeng He, Jiaxuan Wang, Jia Deng

Computer Science and Engineering, University of Michigan, Ann Arbor

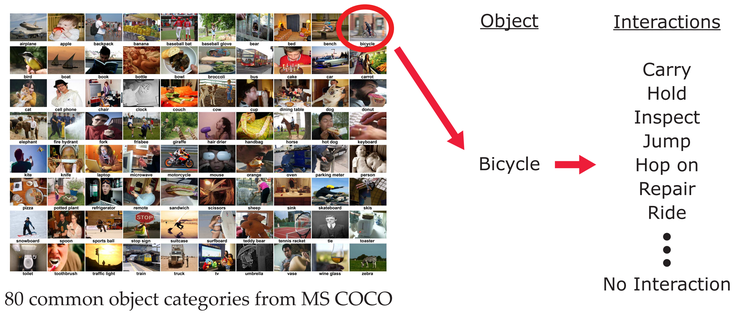
Humans Interacting with Common Objects (HICO)

What is HICO?

A new benchmark for recognizing human-object interactions (HOI).

- 47,774 images in total
- 600 HOI categories over 117 common actions on 80 common objects
- 6.5 distinct interactions per object category on average

1. A *diverse* set of interactions with common object categories.



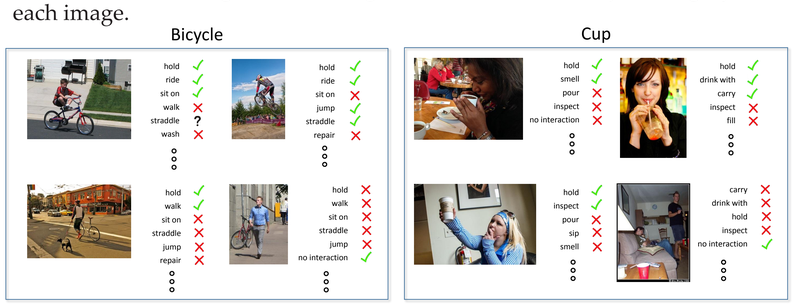
80 common object categories from MS COCO

2. A list of *well-defined, sense-based* HOI categories

Interactions with bicycles

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	

3. An *exhaustive* labeling of *co-occurring* interactions with an object category in each image.



Why HICO?

Existing datasets have limited diversity of interactions with each individual object category. HOI recognition cannot be properly evaluated because a system can "cheat" by simply recognizing the objects.

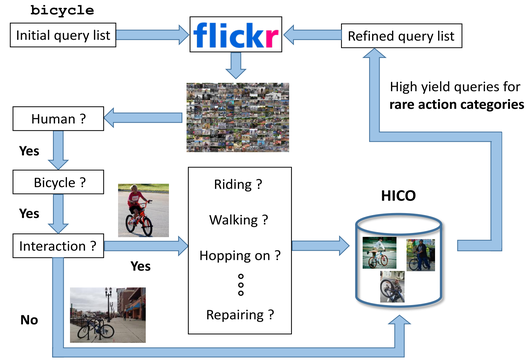
For example, we can recognize "riding a bike" by simply recognizing "bike" in the HMDB dataset.

Object → Interactions

Bicycle → Ride

Constructing HICO

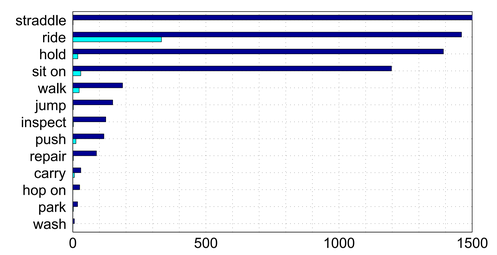
1. Mine common interactions from MS COCO captions and Google N-Gram.
2. Collect and annotate images.



How HICO Differs from Other Datasets

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

HICO v.s. MS COCO on bicycle interactions



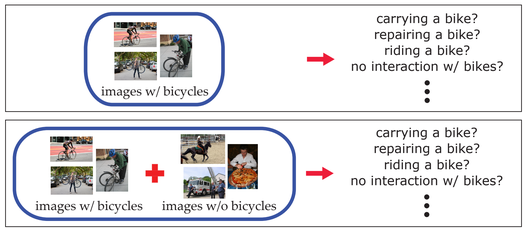
HICO v.s. MPII Human Pose

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

Benchmarking Representative Approaches

Evaluation Setup

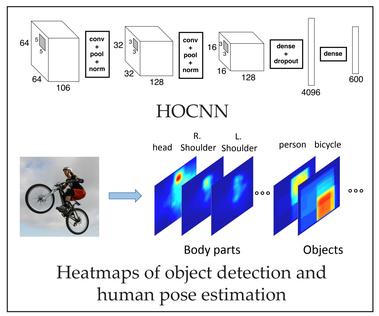
We consider two different settings:

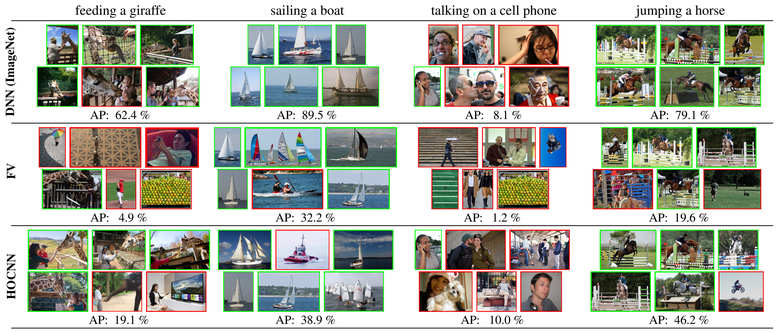


Results

Mean average precision (mAP) on all 600 classes

	mAP	mAP (KO)
Random	0.57	33.37
RandomForest [36]	7.30	38.15
FisherVector [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





Using Semantic Knowledge

1. Knowledge on *Composition*

Classify each HOI by combining outpus of three types of classifiers

VO (e.g. feeding a zebra)

V (e.g. feed)

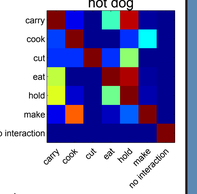
O (e.g. zebra)

feeding a zebra

2. Knowledge on *Coocurrences*

Classify each HOI by considering the classification output of cooccuring HOIs

E.g. classifying "eating a hot dog" by considering the classification output of "holding a hot dog"



mAP in the **default** setting (top) and the **Known Object** setting (bottom)

F: on all 600 HOI classes R: on 167 rare classes (those with less than 5 positive training samples)

	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

Key observation: Semantic knowledge can significantly improve recognition for uncommon categories (R).

References

[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.

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

Dataset and Code

http://www.umich.edu/~ywchao/hico/

Acknowledgement

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