Describing Common Human Visual Actions in Images

Matteo Ruggero Ronchi
http://vision.caltech.edu/~mronchi/
Pietro Perona
perona@caltech.edu

Computational Vision Lab
California Institute of Technology
Pasadena, CA, USA

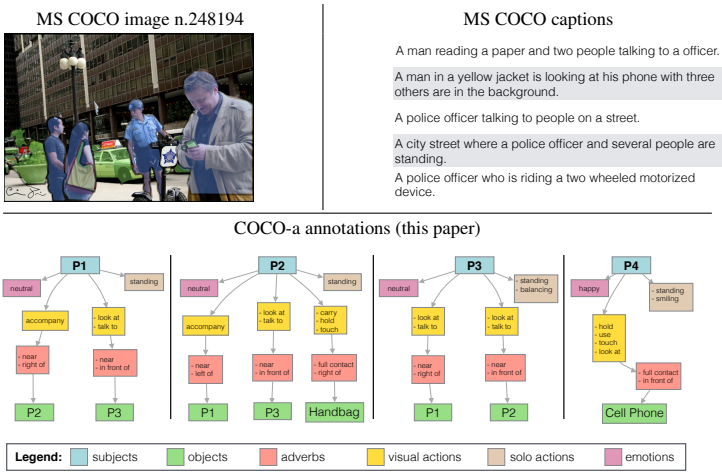

Figure 1: **COCO-a annotations**. (Top) MS COCO image and corresponding captions. (Bottom) COCO-a annotations. Each person (denoted by P1–P4, left to right in the image) is in turn a subject (blue) and an object (green). Annotations are organized by subject. Each subject and each subject-object pair is associated to states and actions. Each action is associated to one of the 140 visual actions in our dataset.

Which common human actions and interactions are recognizable in monocular still images? Which involve objects and/or other people? How many is a person performing at a time? We address these questions by exploring the actions and interactions that are detectable in the images of the MS COCO dataset. We make two main contributions. First, a list of 140 common 'visual actions', obtained by analyzing the largest on-line verb lexicon currently available for English (VerbNet [4]) and human sentences used to describe images in MS COCO. Second, a complete set of annotations for those 'visual actions', composed of subject-object and associated verb, which we call COCO-a (a for 'actions'). COCO-a is larger than existing action datasets in terms of number instances of actions, and is unique because it is data-driven, rather than experimenter-biased. Other unique features are that it is exhaustive, and that all subjects and objects are localized. A statistical analysis of the accuracy of our annotations and of each action, interaction and subject-object combination is provided.

In order to detect actions alongside objects the relationships between those objects needs to be discovered. For each action the roles of 'subject' (active agent) and 'object' (passive - whether thing or person) have to be identified. This information may be expressed as a 'semantic network' [5], which is the first useful output of a vision system for scene understanding.

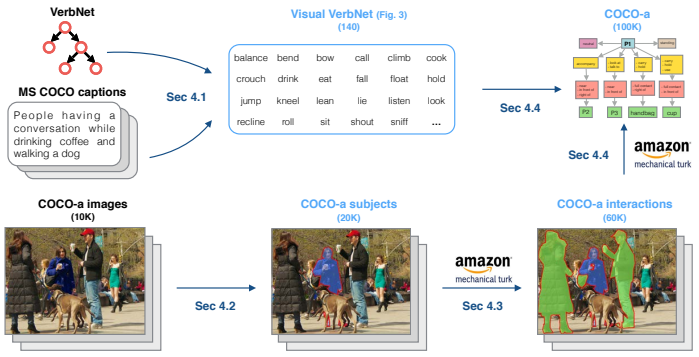

Figure 2: **Steps in the collection of COCO-a.** From VerbNet and MS COCO captions we extracted a list of visual actions. All the persons that are annotated in the MS COCO images were considered as potential 'subjects' of actions, and Amazon Mechanical Turk workers annotated all the objects they interact with, and assigned the corresponding visual actions. Titles in light blue indicate the components of the dataset.

accompany	chew	exchange	jump	pay	punch	sing	swim
avoid	clap	fall	kick	perch	push	sit	talk
balance	clear	feed	kill	pet	put	**skate**	taste
bend (pose)	climb	fight	kiss	photograph	reach	**ski**	teach
bend (something)	cook	fill	kneel	pinch	read	slap	**throw**
be with	crouch	float	laugh	play	recline	sleep	tickle
bite	cry	fly	lay	play baseball	remove	smile	touch
blow	cut	follow	lean	play basketball	repair	sniff	use
bow	dance	get	lick	play frisbee	ride	snowboard	walk
break	devour	give	lie	play instrument	roll	spill	wash
brush	dine	groan	lift	play soccer	row	spray	wear
build	disassemble	groom	**light**	play tennis	run	spread	whistle
bump	draw	hang	listen	poke	sail	squat	wink
call	dress	help	look	pose	separate	squeeze	write
caress	drink	hit	massage	pour	shake hands	stand	
carry	drive	hold	meet	precede	shout	steal	
catch	drop	hug	mix	prepare	show	straddle	
chase	eat	hunt	**paint**	pull	signal	surf	

Figure 3: **Visual VerbNet (VVN).** (Top-Left) The list of 140 visual actions that constitute VVN – bold ones added after the comparison with MS COCO captions. Of the total 2321 verbs in MS COCO captions, there is 60% overlap with the 66 verbs in VVN with > 500 occurrences.

Our goal is to collect an unbiased dataset with a large amount of meaningful and detectable interactions involving human agents as subjects. We put together a process, exemplified in Fig. 2, consisting of four steps: **(1)** Obtain the list of common visual actions that are observed in everyday images, by a combined analysis of VerbNet and MS COCO captions. Our list, which we call Visual VerbNet (Fig. 3) attempts to include all actions that are visually discriminable. It avoids verb synonyms, actions that are specific to particular domains, and fine-grained actions. Unlike previous work, Visual VerbNet is not the result of experimenter's idiosyncratic choices; rather, it is derived from linguistic analysis (VerbNet) and an existing large dataset of descriptions of everyday scenes (MS COCO captions). **(2)** Identify who is carrying out actions (the subjects), as all the people in an image whose pixel area is larger than 1600 pixels. All the people, regardless of size, are still considered as possible objects of an interaction. **(3)** For each subject identify the objects that he/she is interacting with, based on the agreement of 3 out of 5 Amazon Mechanical Turk annotators asked to evaluate each image. **(4)** For each subject-object pair (and each single agent) label all the possible actions and interactions involving that pair, along with high level visual cues such as emotion and posture, spatial relationship and distance.

Our novel dataset, COCO-a, consists of the 'semantic networks' extracted, by following the above process, from 10,000 MS COCO images – examples of the obtained annotations are shown in Fig. 1. MS COCO images are representative of a wide variety of scenes and situations; 81 common objects are annotated in all images with pixel precision segmentations. A key aspect of our annotations is that they provide a complete description of an image (unlike captions which typically focus only on a limited part of it) and unambiguous, as they are based on Visual VerbNet, aiming to eliminate natural language ambiguities and non-visual actions.

We hope that our dataset will provide researchers with a starting point for conceptualizing about actions in images: which representations are most suitable, which algorithms should be used. We also hope that it will provide an ambitious benchmark on which to train and test algorithms. Amongst applications that are enabled by this dataset are building visual Q&A systems [1, 2], more sophisticated image retrieval systems [3], and automated analysis of actions in images of social media.

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: visual question answering. *CoRR*, abs/1505.00468, 2015. URL http://arxiv.org/abs/1505.00468.

[2] D. Geman, S. Geman, N. Hallonquist, and L. Younes. A visual turing test for computer vision system. *Proceedings of the National Academy of Sciences (PNAS)*, 2015.

[3] J. Johnson, R. Krishna, M Stark, Li-Jia Li, D.A. Shamma, M.S. Bernstein, and L. Fei-Fei. Image retrieval using scene graphs. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, 2015.

[4] Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. A large-scale classification of english verbs. *Language Resources and Evaluation*, 42(1):21–40, 2008.

[5] Stuart Russell and Peter Norvig. *Artificial intelligence, a modern approach*. Prentice-Hall, Egnlewood Cliffs, 1995.