

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

Abstract

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) 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.

Appendix Overview

In the appendix we provide:

- (I) Statistics on the type of images in COCO-a.
- (II) Complete list of adverbs and visual actions.
- (III) Complete list of the objects of interactions and occurrence count.
- (IV) Complete list of the visual actions and occurrence count.
- (V) User interface used to collect the interactions in the COCO-a dataset.
- (VI) User interface used to collect the visual actions in the COCO-a dataset.

Appendix I: Unbiased Nature of COCO-a

We show in Figure 9 the unbiased nature of the images contained in our dataset. Different actions usually occur in different environments, so in order to balance the content of our dataset we selected an approximately equal number images of three types of scenes: sports, outdoors and indoors. We also selected images of various complexity, containing single subjects, small groups (2-4 subjects) and crowds (>4 subjects). The image selection process consists of the following two steps: (1) categorize all the images containing people in MS COCO based on the types of objects they contain; (2) randomly sample images in about equal percentages from all categories and complexities.

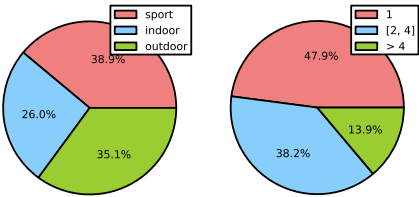


Figure 9: **Scene and subjects distributions.** (Left) The distribution of the type of scenes contained in the dataset. (Right) The distribution of the number of subjects appearing in each image.

Appendix II: Visual Actions and Adverbs by Category

In order to reduce the possibility of annotators using a term instead of another in the data collection interface, we organized visual actions into 8 groups – ‘*posture/motion*’, ‘*solo actions*’, ‘*contact actions*’, ‘*actions with objects*’, ‘*social actions*’, ‘*nutrition actions*’, ‘*communication actions*’, ‘*perception actions*’. This was based on two simple rules: (a) actions in the same group share some important property, e.g. being performed solo, with objects, with people, or indifferently with people and objects, or being an action of posture; (b) actions in the same group tend to be mutually exclusive, e.g. a person can be drinking or eating at a certain moment, not both. Furthermore, we included in our study 3 ‘adverb’ categories: ‘*emotion*’ of the subject, ‘*location*’ and ‘*relative distance*’ of object with respect to the subject. Tables 2 and 3 contain a break down of the visual actions and adverbs into the categories that were presented to the Amazon Mechanical Turk workers.

Adverbs		
Emotion (6)	Relative Location (6)	Relative Distance (4)
anger	above	far
disgust	behind	full contact
fear	below	light contact
happiness	in front	near
sadness	left	
surprise	right	

Table 2: **Adverbs ordered by category.** The complete list of high level visual cues collected, describing the subjects (emotion) and localization of the interaction (relative location and distance).

Visual Actions								
Posture / Motion (23)				Communication (6)		Contact (22)		
balance	hang	run		call		avoid	massage	
bend	jump	sit		shout		bit	pet	
bow	kneel	squat		signal		bump	pinch	
climb	lean	stand		talk		caress	poke	
crouch	lie	straddle		whistle		hit	pull	
fall	perch	swim		wink		hold	punch	
float	recline	walk				hug	push	
fly	roll					kick	reach	
						kiss	slap	
						lick	squeeze	
						lift	tickle	
Social (24*)				Perception (5)		Nutrition (7)		
accompany	give	play baseball		listen		chew		
be with	groom	play basketball		look		cook		
chase	help	play frisbee		sniff		devour		
dance	hunt	play soccer		taste		drink		
dine	kill	play tennis		touch		eat		
dress	meet	precede				prepare		
feed	pay					spread		
fight	shake hands							
follow	teach							
Solo (24*)				With objects (34)				
blow	play soccer			bend	fill	separate		
clap	play tennis			break	get	show		
cry	play instrument			brush	lay	spill		
draw	pose			build	light	spray		
groan	sing			carry	mix	steal		
laugh	sleep			catch	pour	put		
paint	smile			clear	read	throw		
photograph	write			cut	remove	use		
play	skate			disassemble	repair	wash		
play baseball	ski			drive	ride	wear		
play basketball	snowboard			drop	row			
play frisbee	surf			exchange	sail			

Table 3: **Visual actions ordered by category.** The complete list of visual actions contained in Visual VerbNet. Visual actions in one category are usually mutually exclusive, visual actions of different categories may co-occur. (*) There are five visual actions (*play baseball*, *play basketball*, *play frisbee*, *play soccer*, *play tennis*) that are considered both ‘social’ and ‘solo’ types of actions.

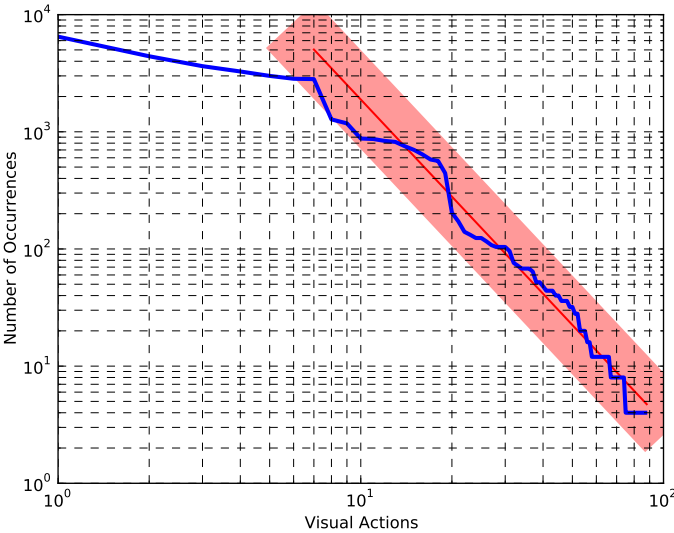


Figure 12: **Visual actions heavy tail analysis.** The plot in log-log scale of the list of visual actions against the number of occurrences.

Appendix V: Interactions User Interface

In Figure 13 we show the AMT interface developed to collect interaction annotations from images. Each worker is presented with a series of 10 images, each containing a subject highlighted in blue and asked to (1) flag the subject if it is mostly occluded or invisible; (2) if the subject is sufficiently visible, click on all the objects he/she is interacting with. The interface provides feedback to the annotator by highlighting in white all the annotated objects when the mouse is hovered over the image, and selecting in green the objects once they are clicked. Annotators can remove annotations by either clicking on the object segmentation on the image a second time or using the appropriate button in the annotation panel. We included a comments text box to obtain specific feedback workers on each image.

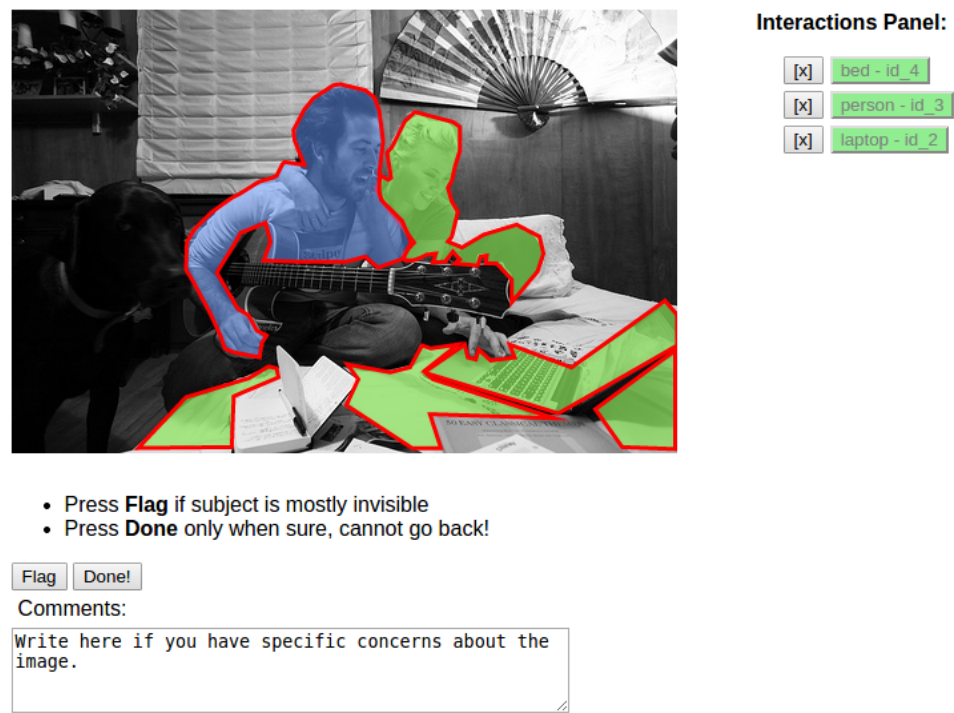
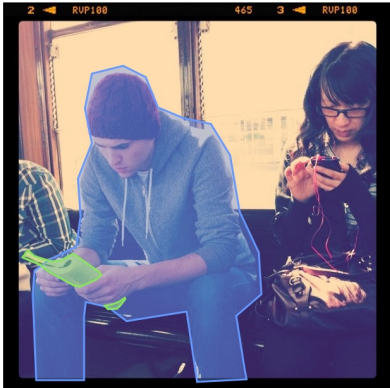


Figure 13: **Interactions GUI.** In this image the blue subject is interacting with another person, the bed and the laptop.

Appendix VI: Visual Actions User Interface

In Figures 14, 15, 16, 17 and 18 we show the sequences of steps required in the AMT interface developed to collect visual action annotations. We collect visual actions for all the interactions obtained from the previously shown GUI having an agreement of 3 out of 5 workers, as explained in more details in Section 4.3 of the main paper. Each worker is presented with a single image containing a subject (highlighted in blue) and an object (highlighted in green) and asked to go through 8 panels, one for each category of visual actions, and select all the visual actions that apply to the visualized interaction. Annotators can skip a category if no visual action applies (i.e. nutrition visual actions only apply for food items). As they proceed through the 8 panels workers have the chance to visualize all the annotations that are being provided for the specific interaction, which helps avoid ambiguous annotations. Depending on the object involved in the interaction some panels might not be shown (i.e. the *communication* panel is not shown when the object of interaction is inanimate, as well as the *nutrition* panel is not shown when the object of interaction is another person).



Step 1: Flag the interaction if subject is occluded

Press **Flag** if:

- the blue **subject** is occluded or invisible in such a way that you cannot determine his actions.
- the green **object** is occluded in such a way that you cannot determine the blue **subject**'s actions with it.
- there are multiple blue **subjects** or green **objects**.

Don't Flag

Flag

Figure 14: Visual Actions GUI. (I/V)

Step 2: Provide Relative Location

Relative Location: Where is the green **object** with respect to the blue **subject**?

Answer as if you were the blue subject and had to give the position of the green object with respect to you.

Press **Next** if none apply.

A: **B:** **I:** **L:** **R:**

Annotations:

Step 3: Provide Distance of Interaction

Distance of Interaction: What is the distance between the blue **subject** and the green **object**?

- **Full contact:** if subject is sorrounding or holding the object
- **Light contact:** if subject is touching or patting the object
- **Near to:** if subject is within a few feet from the object
- **Far from:** if subject is more than a few feet from the object

F: **L:** **N:**

Annotations:

Figure 15: **Visual Actions GUI.** (II/V)

Step 4: Provide Senses used in Interaction

Perception: Which of his senses is the blue **subject** using to interact with the green **object** (if recognizable / any)?

Press **Next** if none apply.

L: **S:** **T:**

Annotations:

☐

☐

☐

☐

Step 5: Provide Nutrition Visual Actions (none in this case)

Nutrition: Which of these nutrition actions is the blue **subject** involved with the green **object** (if applies)?

C: **D:** **E:** **P:** **S:**

Figure 16: **Visual Actions GUI.** (III/V)

Step 6: Provide Contact Visual Actions (free-typing is allowed)

Subject-Object Interactions (1): How is the blue **subject** interacting with the green **object**?

A:

is avoiding

B:

is biting

is bumping

C:

is caressing

H:

is hitting

is holding

is hugging

K:

is kicking

is kissing

L:

is licking

is lifting

M:

is massaging

P:

is petting

is pinching

is poking

is pulling

is punching

is pushing

R:

is reaching

S:

is slapping

is squeezing

T:

is tickling

Add custom annotation:

Use the box below to add a custom annotation only if you are not able to describe the **Subject-Object Interactions (1)** with any of the predicates above.

Person

Add one if you think it's missing.

book

Add

Annotations:

[x]

book

is in front of

Person

[x]

Person

in full contact with

book

[x]

Person

is looking at

book

[x]

Person

is touching

book

[x]

Person

is holding

book

Figure 17: Visual Actions GUI. (IV/V)

Step 7: Provide Object Visual Actions (free-typing is allowed)

Subject-Object Interactions (2): How is the blue **subject** interacting with the green **object**?

B: **C:**

D: **E:** **F:** **G:** **L:**

M: **P:** **R:**

S: **T:**

U: **W:**

Add custom annotation:

Use the box below to add a custom annotation only if you are not able to describe the **Subject-Object Interactions (2)** with any of the predicates above.

Annotations:

Figure 18: Visual Actions GUI. (V/V)