FS-COCO: Towards Understanding of Freehand Sketches of Common Objects in Context

Pinaki Nath Chowdhury[1,2] Aneeshaan Sain[1,2] Ayan Kumar Bhunia[1] Tao Xiang[1,2] Yulia Gryaditskaya[1,2,3] Yi-Zhe Song[1,2]

[1]SketchX, CVSSP, University of Surrey, United Kingdom.
[2]iFlyTek-Surrey Joint Research Centre on Artificial Intelligence.
[3]Surrey Institute for People-Centred AI, CVSSP, University of Surrey, United Kingdom.

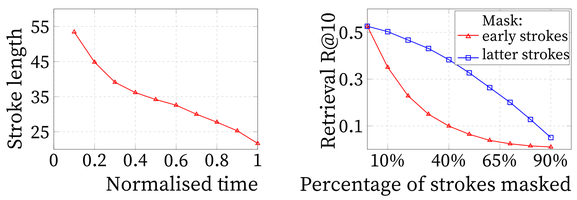
Contributions

- We propose the first dataset of freehand scene sketches and their captions.
- We study the relations between sketches, images and their captions.
- We study for the first time fine-grained freehand scene sketch-based image retrieval.
- We show the limitation of a popular CLIP [1] model on the fine-grained scene sketch-based image retrieval task.
- We introduce a novel hierarchical sketch decoder that exploits temporal stroke order.
- We test machine understanding of sketches by evaluating several state-of-the-art image captioning methods on our sketches.

What have we learned about scene sketches?

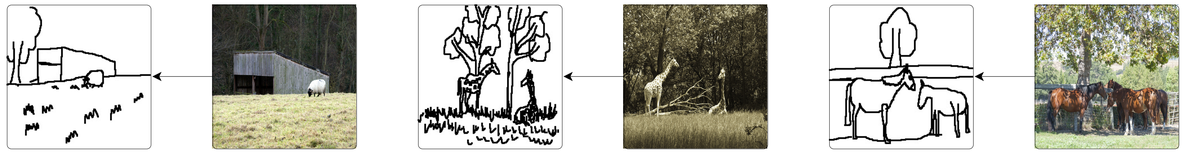
(a) Image captions (b) Sketch captions (c) Statistics

Word clouds show frequently occurring words in sketch and image captions. By studying captions we analyze what information is contained in sketches. Sketch captions focus more on objects (i.e., nouns like "horse") and their actions (i.e., verbs like "standing"), instead of focusing on attributes (i.e., adjectives like "a brown horse").



Humans follow coarse-to-fine sketching strategy (left), drawing most salient for retrieval task strokes first (right).

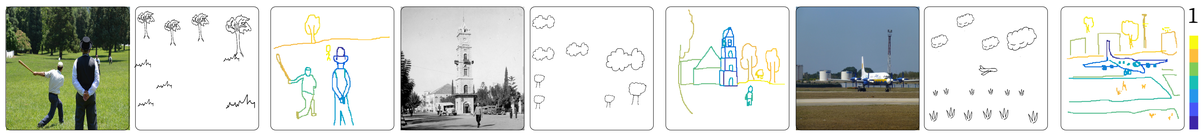
FS-COCO Dataset



A sheep is eating the grass | A white sheep standing on a dry grass field | One giraffe standing on a grass and one giraffe sitting on a grass | The giraffes are looking at each other in the field | Horses are standing near a fence and tree | A couple of brown horses standing next to trees

We introduce the FS-COCO (Freehand Sketches of Common Objects in COntext) dataset. It is the first dataset of 10,000 unique freehand scene sketches, drawn by 100 participants. For every sketch we collect its caption. The textual description makes the annotator who created the sketch, eliminating the noise (ambiguity) due to sketch interpretation.

Comparison with other scene-sketch datasets

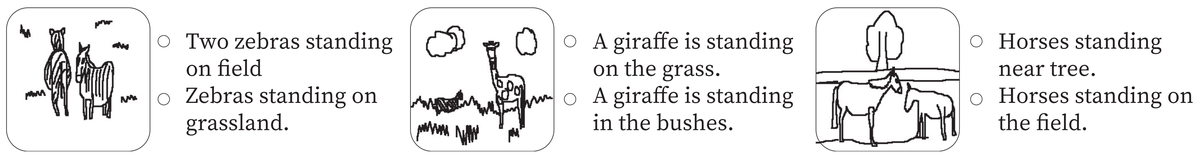


Compared to previous datasets `SketchyCOCO` [2] and `SketchyScene` [3], our freehand scene sketches contain abstraction at both object and scene level.

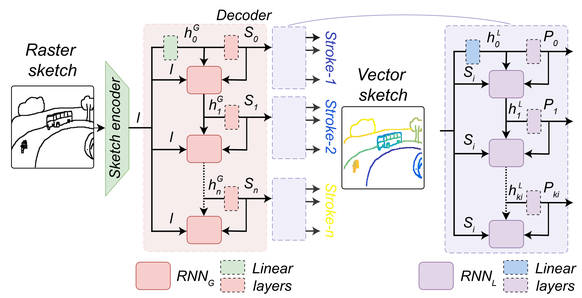
Dataset	# photos	#categories	# categories per sketch Mean	Std	Min	Max	# sketches per category Mean	Std	Min	Max
`SketchyScene`	7,264	45	7.88	1.96	4	20	1079.76	1447.47	31	5723
`SketchyCOCO`	14,081	17	3.33	0.9	2	7	1932.41	3493.01	33	9761
`SketchyScene FG`	2,724	45	7.71	1.88	4	20	394.51	540.30	3	2154
`SketchyCOCO FG`	1,225	17	3.28	0.89	2	6	164.71	297.79	5	824
FS-COCO (c)	10,000	92	1.37	0.57	1	5	99.42	172.88	1	866
FS-COCO (l)	10,000	150	7.17	3.27	1	25	413.18	973.59	1	6789

Comparison of scene sketch datasets based on the distribution of categories in sketch-image pairs. 'FG' denotes subsets of datasets that are recommended for Fine-Grained tasks. e_l/e_c denotes estimates based on semantic segmentation labels in images, and the occurrence of a word in a sketch caption, respectively.

Sketch understating via automatic sketch captioning



Hierarchical Decoder



We propose a novel 2-layered LSTM decoder for pre-training a sketch encoder – by solving a "pretext" task of converting a rasterized sketch to its vectorized format. Existing single-layer Recurrent Neural Networks (RNN) fail on scene sketches since they can only model up to around 200 stroke points. Our scene sketches contain more than 3000 stroke points. However, we build our decoder based on the observation that, on average, scene sketches consist of only 74.3 strokes, with each stroke containing around 41.1 stroke points.

The role of pre-training with H-Decode in retrieval:

	Baseline		H-Decoder	
Method	R@1	R@10	R@1	R@10
Siam.-VGG16	23.3	52.6	**24.1**	**54.3**
CLIP*	5.5	26.5	5.7	27.1

More Information

https://fscoco.github.io

[1] Radford et al.: Learning transferable visual models from natural language supervision, 2021
[2] Gao et al.: SketchyCOCO: Image generation from freehand scene sketches, 2020
[3] Zou et al.: SketchyScene: Richly-annotated scene sketches, 2018