Sketch-a-Net that Beats Humans

Qian Yu*
q.yu@qmul.ac.uk
Yongxin Yang*
yongxin.yang@qmul.ac.uk
Yi-Zhe Song
yizhe.song@qmul.ac.uk
Tao Xiang
t.xiang@qmul.ac.uk
Timothy M. Hospedales
t.hospedales@qmul.ac.uk

School of Electronic Engineering and Computer Science
Queen Mary, University of London
London, E1 4NS
United Kingdom

Sketches are very intuitive to humans and have long been used as an effective communicative tool. With the proliferation of touchscreens, sketching has become a much easier undertaking for many – we can sketch on phones, tablets and even watches. However, recognising free-hand sketches (e.g. asking a person to draw a car without any instance of car as reference) is an extremely challenging task. This is due to a number of reasons: (i) sketches are highly iconic and abstract, e.g., human figures can be depicted as stickmen; (ii) due to the free-hand nature, the same object can be drawn with hugely varied levels of detail/abstraction, e.g., a human figure sketch can be either a stickman or a portrait with fine details depending on the drawer; (iii) sketches lack visual cues, i.e., they consist of black and white lines instead of coloured pixels. A recent large-scale study on 20,000 free-hand sketches across 250 categories of daily objects puts human sketch recognition accuracy at 73.1% [2], suggesting that the task is challenging even for humans.

Prior work on sketch recognition generally follows the conventional image classification paradigm, that is, extracting hand-crafted features from sketch images followed by feeding them to a classifier. Most hand-crafted features traditionally used for photos (such as HOG, SIFT and shape context) have been employed, which are often coupled with Bag-of-Words (BoW) to yield a final feature representations that can then be classified. However, existing hand-crafted features designed for photos do not account for the unique abstract and sparse nature of sketches. Furthermore, they ignore a key unique characteristics of sketches, that is, a sketch is essentially an ordered list of strokes; they are thus sequential in nature (See Fig 1). In contrast with photos that consist of pixels sampled all at once, a sketch is the result of an online drawing process. It had long been recognised in psychology that such sequential ordering is a strong cue in human sketch recognition, a phenomenon that is also confirmed by recent studies in the computer vision literature [7]. However, none of the existing approaches attempted to embed sequential ordering of strokes in the recognition pipeline even though that information is readily available.

Figure 1: Illustration of stroke ordering in sketching with the Alarm Clock category. Each sketch is split into three parts according to stroke ordering.

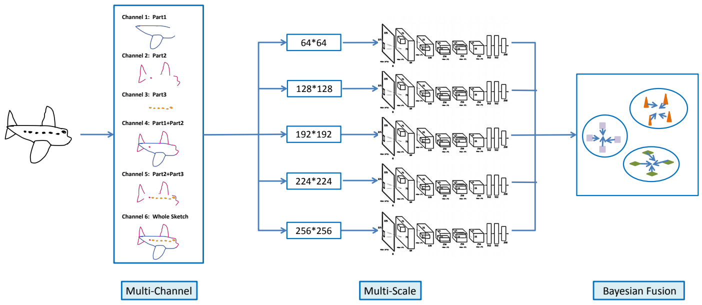

Figure 2: Illustration of our overall framework.

In this paper, we propose a novel deep neural network (DNN), Sketch-a-Net (See Fig 2), for free-hand sketch recognition, which is specifically designed to accommodate the unique characteristics of sketches including multiple levels of abstraction and being sequential in nature. Our contributions are summarised as follows: (i) for the first time, a representation learning model based on DNN is presented for sketch recognition in place of the conventional hand-crafted feature based sketch representations (Details are listed in Table 1); (ii) we demonstrate how sequential ordering information in sketches can be embedded into the DNN architecture and in turn improve sketch recognition performance; (iii) we propose a multi-scale network ensemble that fuses networks learned at different scales together via joint Bayesian fusion [1] to address the variability of levels of abstraction in sketches. Extensive experiments on the largest hand-free sketch benchmark dataset, the TU-Berlin sketch dataset [2], show that our model significantly outperforms existing approaches and can even beat humans by 1.8% at sketch recognition (See Table 2).

Index	Layer	Type	Filter Size	Filter Num	Stride	Pad	Output Size
0		Input	-	-	-	-	225×225
1	L1	Conv	15×15	64	3	0	71×71
2		ReLU	-	-	-	-	71×71
3		Maxpool	3×3	-	2	0	35×35
4	L2	Conv	5×5	128	1	0	31×31
5		ReLU	-	-	-	-	31×31
6		Maxpool	3×3	-	2	0	15×15
7	L3	Conv	3×3	256	1	1	15×15
8		ReLU	-	-	-	-	15×15
9	L4	Conv	3×3	256	1	1	15×15
10		ReLU	-	-	-	-	15×15
11	L5	Conv	3×3	256	1	1	15×15
12		ReLU	-	-	-	-	15×15
13		Maxpool	3×3	-	2	0	7×7
14	L6	Conv(=FC)	7×7	512	1	0	1×1
15		ReLU	-	-	-	-	1×1
16		Dropout (0.50)	-	-	-	-	1×1
17	L7	Conv(=FC)	1×1	512	1	0	1×1
18		ReLU	-	-	-	-	1×1
19		Dropout (0.50)	-	-	-	-	1×1
20	L8	Conv(=FC)	1×1	250	1	0	1×1

Table 1: The architecture of Sketch-a-Net.

HOG-SVM [2]	Ensemble [5]	MKL-SVM [6]	FV-SP [7]	
56%	61.5%	65.8%	68.9	
AlexNet-SVM [3]	AlexNet-Sketch [3]	LeNet [4]	Sketch-a-Net	Human [2]
67.1%	68.6%	55.2%	**74.9%**	73.1%

Table 2: Comparison with state of the art results on sketch recognition

* These authors contributed equally to this work

[1] D. Chen, X. Cao, L. Wang, F. Wen, and J. Sun. Bayesian face revisited: A joint formulation. In *ECCV*, 2012.

[2] M. Eitz, J. Hays, and M. Alexa. How do humans sketch objects? In *SIGGRAPH*, 2012.

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

[4] Y. LeCun, L. Bottou, G. B. Orr, and K. Müller. Efficient backprop. *Neural networks: Tricks of the trade*, pages 9–48, 2012.

[5] Y. Li, Y. Song, and S. Gong. Sketch recognition by ensemble matching of structured features. In *BMVC*, 2013.

[6] Y. Li, T. M. Hospedales, Y. Song, and S. Gong. Free-hand sketch recognition by multi-kernel feature learning. *CVIU*, 2015.

[7] R. G. Schneider and T. Tuytelaars. Sketch classification and classification-driven analysis using fisher vectors. In *SIGGRAPH Asia*, 2014.