Human's Scene Sketch Understanding

Yuxiang Ye
Texas State University
y_y13@txstate.edu

Yijuan Lu
Texas State University
lu@txstate.edu

Hao Jiang
Boston College
hjiang@cs.bc.edu

ABSTRACT

Human's sketch understanding is important. It has many applications in human computer interaction, multimedia, and computer vision. Recognizing human sketches is also challenging. Previous methods focus on single-object sketch recognition. Understanding human's scene sketch that involves multiple objects and their complex interactions has not been explored. In this paper, we tackle this new problem. We create the first scene sketch dataset "Scene250" and propose a deep learning method to understand human scene sketches. We propose "Scene-Net", a new deep convolutional neural network (CNN) structure, based on which we build a novel scene sketch recognition system. Our system has been tested on the collected scene sketch dataset and compared with other state-of-the-art CNNs and sketch recognition approaches. Our experimental results demonstrate that our method achieves the state of art.

Categories and Subject Descriptors

I.5 [**Pattern Recognition**]: Miscellaneous

Keywords

Sketch understanding; scene sketch; deep learning

1. MOTIVATION

Since prehistoric times, sketching has been a unique way to visually render human's mind. In nowadays, with the increasing popularity of devices with touch screens (e.g., touch pads and smart phones), sketching has become one of the most natural means of human-computer interaction. Sketching on the a scene of a smart phone has been used to provide the input to an image retrieval system. It can be used as a frontal end for generating full 3D models. Sketching also provides an attractive interface for children, especially pre-school kids, to interact with computers. Teaching computers to understand hand-drawn sketches is thus valuable. Sketch understanding has received more and more interests recently.

Sketch understanding is also quite challenging. The difficulty is mainly because a sketch only contains lines and lacks color, texture, and enough visual cues compared to color images. Human sketches also often depict complex high-level concept. Current sketch-related methods mainly focus on recognizing sketches that contain only one single object [1, 9] in each image. However, these single-object sketches are very different from the sketches that people usually draw. In a human drawn sketch, we often see complete scenes or even a depicted story (Fig. 1). Without the information of object regions and segmentations, understanding a scene sketch with multiple objects is more challenging than recognizing single-object sketch. Another factor that complicates scene sketch recognition is the possible variation of rotation, scale and view point as shown in Fig. 2. There is also an uncertainly about the drawing style. One thousand people may sketch the same scene in one thousand different ways. When sketching the same scene, different people tend to choose different set of objects (Fig. 2: Desert). The sketches of "River" can be either several strokes to depict the river shape or a portrait with fine details of plants along the river (Fig. 2: River). So far, there is a lack of a comprehensive study of how human's scene sketches can be well understood by computers.

In this paper, we study scene sketch understanding. We create the first scene sketch dataset Scene250 and explore deep learning approach to scene sketch understanding. We propose a new deep CNN structure and build a novel scene sketch understanding system based on this model. The performance of our system has been tested on the collected scene sketch dataset and compared with other state-of-the-art sketch recognition approaches. Our methods give much better results than these previous methods.

To our best knowledge, this work is the first attempt to explore scene sketch understanding and to construct a deep learning CNN to solve the problem. Our main contributions introduced in this paper are highlighted as follows:

- The first scene sketch dataset is created and open to public. It contains 250 scene sketches in 10 categories. The sketches include common indoor and outdoor scenes.
- A new scene sketch deep CNN structure is proposed to target the scene sketch recognition task.
- Comprehensive experiments have been conducted to evaluate the state-of-the-art sketch recognition approaches on human's scene sketch recognition.

ICMR '16, June 6–9, 2016, New York, NY, USA.

DOI: http://dx.doi.org/10.1145/2911996.2912067

- Our proposed scene sketch recognition method provides enabling technique for different sketch-based applications.

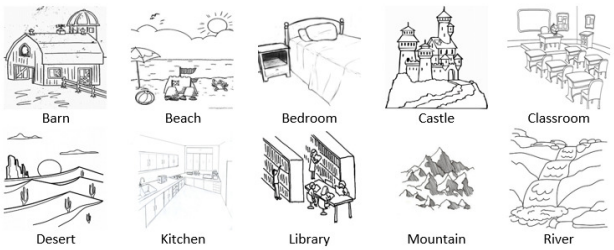

Figure 1: Example scene sketches of our Scene250 dataset (one example per category)

2. RELATED WORK

In scene sketch recognition, we classify the input into different scene categories. In the following, we first review current public domain sketch datasets, and then we discuss previous sketch recognition approaches.

(1) Dataset. Early sketch datasets such as the artistic drawing dataset [7] and the structure sketch dataset [6] are either small or restricted to a specific domain. The more recent TU Berlin dataset [1] is larger. It contains 20,000 single-object sketches in 250 daily object categories. In contrast to the TU Berlin dataset, in this paper, we deal with scene sketches that involve multiple objects and complex interaction among these objects. We have constructed a new dataset Scene250 that contains 250 sketches in ten categories.

(2) Sketch recognition. Early sketch recognition methods [2, 8] have been used to enhance the human computer interaction experience. They use direct inputs from drawings on touch screens or mouses. Such inputs are often noise free. Nowadays, sketch recognition methods have been developed to handle line drawings in noisy scanned or photographed images. Different hand-crafted features, such as stroke length, stroke order and stroke orientation have been used in sketch recognition. Eitz et al. [4] proposed to use sHOG and bag-of-word methods for sketch understanding. Even though this method gave a promising correct classification rate of 56% on the TU Berlin sketch dataset, designing effective hand-crafted features is a challenging task. It requires extensive knowledge on drawing. And, there is also no guide line about how to construct the optimal features. More flexible sketch recognition approach needs to be designed.

Deep learning, which tackles learning the features and classifiers simultaneously, is a promising framework to tacking the problem. Convolution Neural Networks (CNNs) have shown remarkable results in many vision tasks of different domains. With the introduction of rectifier linear (ReLU) [5], max-pooling, local response normalization (LRN) [4], and dropout regularization units [3], CNNs become less likely to overfit. They generalize well on unseen data. Yu et al. [9] designed a sketch-oriented deep CNN model "Sketch-a-Net" for sketch recognition task and achieved the accuracy of 74.9% on TU Berlin dataset.

Most previous works only focus on single-object sketch recognition. However, in realistic world, people usually draw a sketch with multiple objects, i.e., a scene sketch, rather than a simple single-object sketch. Compared with single-object sketch recognition, scene sketch understanding is more challenging. Some examples are shown in Fig. 2.

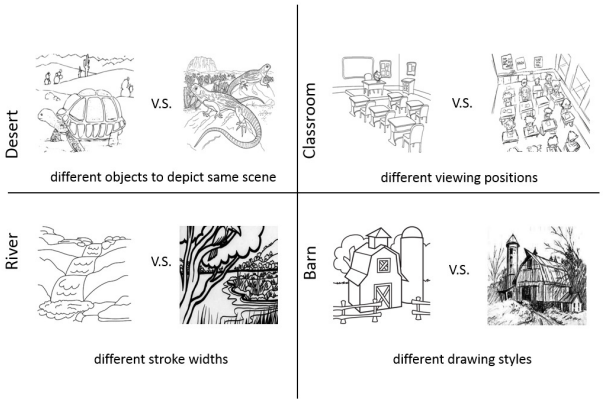

Figure 2: Examples to demonstrate scene sketch recognition challenges

3. METHOD

3.1 A deep CNN model

We propose a novel scene sketch oriented deep CNN model "Scene-Net". Its architecture is illustrated in Fig. 3, while more detailed parameters can be found in Table 1. As shown in Fig. 3, our Scene-Net CNN contains five convolutional layers (L1∼L5) followed by three fully-connected layers (L6∼L8). Each layer has a ReLU unit except for the layer L8. LRN units are appended to the layers L1 and L2 while max pooling units are appended to the layers L1, L2 and L5. We apply dropout units to the first two fully connected layers (L6 and L7). The third fully connected layer (L8), which is appended by a softmax loss layered, is the last layer of our CNN model. The output size of the last layer is 10, which corresponds to the 10 scene categories.

3.2 Smaller Input Size

Setting the input size of a CNN is critical, since all the following layers depend on the input layer. Many modern deep CNNs [4, 9] use large input size around 225×225. However, we find that smaller input size is more appropriate for scene sketch recognition. The major reason is that scene sketches usually contain multiple objects within one image. Smaller input size combined with random data augmentation can capture more sophisticated features. To this end, input size of 193×193 is used in our Scene-Net.

3.3 Local Response Normalization

Local Response Normalization (LRN) performs a kind of "lateral inhibition" by normalizing inputs in local regions. In this paper, we denote the size of local region as n, the scaling parameter as α, and the exponent as β in Table 1.

3.4 Dropout

Dropout reduces complex co-adaptations of neurons by randomly setting unit activation to zero. The "dropped out" neurons do not participate in the forward pass and back-propagation. In our Scene-Net, we set the dropout rate as

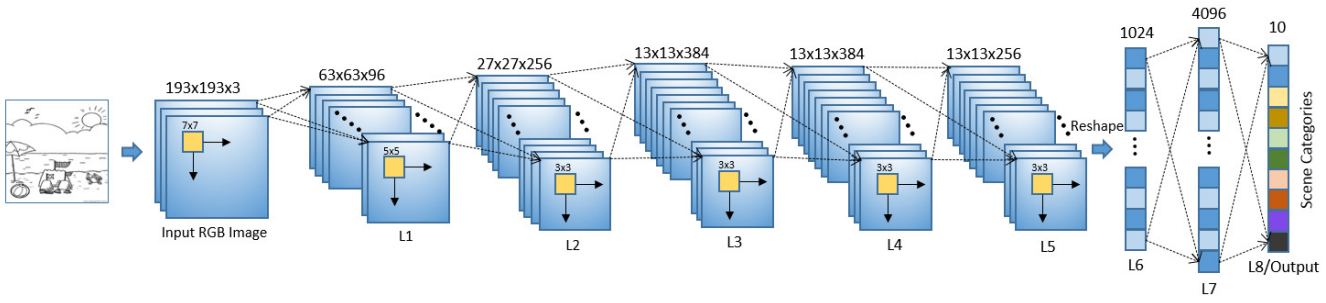

Figure 3: Overall architecture of Scene-Net model

0.5, which is a reasonable approximation. Dropout layers are appended to the first two fully connected layers.

3.5 Fine-tuning

We use fine-tuning to deal with the small training dataset problem. Since our training dataset is very small, directly training the neural network with millions of parameters would make cause the overfit of the trained classier on the small dataset. We there first pre-trained Scene-Net model on the TU Berlin dataset and obtained the initial learning weights for the Scene-Net model. Then, we resumed the training process on the Scene250 dataset to fine tune Scene-Net model for scene sketch recognition task. In our experiments, we find applying fine-tuning significantly increases the recognition accuracy and prevents the overfitting problem.

4. EXPERIMENTS AND DISCUSSIONS

4.1 Scene250 Dataset Collection

We create the scene sketch dataset "Scene250". Scene250 (Fig. 1) contains 250 scene sketches from the Internet. These sketches cover 10 different categories and each category includes 25 sketches. To construct the dataset, we searched for sketches via Google using different key words. We downloaded more than 100 sketches for each category and selected 25 representative ones. The following criteria are used when selecting the categories and scene sketches in each category.

- Completeness. Although the categories of Scene250 does not cover all the scene categories, the selected 10 categories represent the most common ones in the everyday life and contain both indoor and outdoor scenes.
- Unambiguous. Each selected sketches that can be classified into more than one category. For example, we don't select the sketches that can be classified as either "desert" or "river".
- Recognizable. Human observer should be able to recognize these selected sketches from their shape alone without other context information such as text.
- Diversity. Sketches in the same scene category should be diversified. We avoid selecting the sketches that are visually identical in order to guarantee the diversity.

We manually inspected the whole dataset by displaying sketches on a screen. All the scene sketches in the same category were displayed together, which allowed us to identify improper ones easily. We removed those sketches that are in the wrong category (e.g., a mountain scene in the bedroom category) or not subject to our selection constraints. We also rescaled the dataset to contain exactly 25 scene sketches per category yielding the final 250 scene sketches in the Scene250 dataset. Benefited from consistent size of each category, there is no need to correct for bias toward the categories with different size when performing training or testing.

4.2 Image Format

We rescaled all the images to 256×256 pixels and read the images as single channel or multiple channels, which depends on the architecture of CNNs.

4.3 Data Augmentation

We randomly perturb the data to generate a bigger set of training data to prevent overfitting. In our experiments, each image is perturbed randomly to generate 500 different inputs. Each image is randomly shifted in the x or/and y directions by 0-63 pixels. The result is further rotated randomly in 0-360 degrees and then flipped vertically with a 50% chance.

4.4 Comparison

We compared our Scene-Net model with two other deep CNN models: AlexNet [4] and Sketch-a-Net. Although AlexNet is a photo-oriented deep CNN model designed for ImageNet, it gives good results in many different applications (e.g., video, robotics, and bioinformatics). Sketch-a-Net is a sketch-oriented deep CNN model specifically designed for single-object sketch classification. It even beats human recognition accuracy by 1.8% on TU Berlin dataset. Both AlexNet and Sketch-a-Net have five convolutional layers followed by three fully-connected layers (last fully-connected layer is the output layer). Sketch-a-Net uses larger first layer filters (11×11), higher dropout rate (0.55), and overlapped pooling.

We compared our Scene-Net with the above CNN methods using our Scene250 dataset. Following the settings in previous works, we used 2/3 sketches per category for training and 1/3 for testing. In addition, since there are two types of scene sketches in Scene250: indoor scene (i.e., bedroom, classroom, kitchen, and library) and out-door scene (i.e., barn, beach, castle, desert, mountain, and river), we also tested these CNN models on in-door scene and out-door scene recognition individually. The comparison results are listed in Table 2 and 3. Our Scene-Net model not only beats AlexNet and Sketch-a-Net on overall recognition performance but also beats them on both in-door and out-door scene recognition task.

Table 1: Detailed architecture of Scene-Net

Index	Layer	Type	Filter Size	Filter Num	Stride	Pad	Output Size
0		Input	-	-	-	-	193×193
1	L1	Conv	7×7	96	3	0	63×63
2		ReLU	-	-	-	-	63×63
3		LRN($n = 5, \alpha = 10^{-4}, \beta = 0.75$)	-	-	-	-	63×63
4		Maxpool	3×3	-	2	0	31×31
6	L2	Conv	5×5	256	1	0	27×27
7		ReLU	-	-	-	-	27×27
8		LRN($n = 5, \alpha = 10^{-4}, \beta = 0.75$)	-	-	-	-	27×27
9		Maxpool	3×3	-	2	0	13×13
10	L3	Conv	3×3	384	1	1	13×13
11		ReLU	-	-	-	-	13×13
12	L4	Conv	3×3	384	1	1	13×13
13		ReLU	-	-	-	-	13×13
12	L5	Conv	3×3	256	1	1	13×13
13		ReLU	-	-	-	-	13×13
14		Maxpool	3×3	-	2	0	6×6
15	L6	Fully-connected	6×6	1024	1	0	1×1
16		ReLU	-	-	-	-	1×1
17		Dropout(0.5)	-	-	1	0	1×1
18	L7	Fully-connected	1×1	4096	1	0	1×1
19		ReLU	-	-	-	-	1×1
20		Dropout(0.5)	-	-	1	0	1×1
18	L8	Fully-connected	1×1	10	1	0	1×1
19		Softmax loss	-	-	-	-	1×1

4.5 Implementation details

We implemented Scene-Net model using Matlab and the MatConvNet toolbox. All the experiments were executed on a Linux machine with an 8-core 3.50GHz CPU and a GeForce GTX Titan X GPU. The pre-training time for our Scene-Net model on the TU Berlin dataset is approximately 8 hours on GPU, while fine-tuning with Scene250 dataset is about 3 hours on GPU.

Table 2: Scene sketch recognition performance comparison

Scene-Net	AlexNet	Sketch-a-Net
0.6000	0.5250	0.3875

Table 3: In-door/Out-door scene sketch recognition comparison

	Scene-Net	AlexNet	Sketch-a-Net
In-door	0.5000	0.4688	0.2500
Out-door	0.6667	0.5625	0.4792

5. CONCLUSIONS AND FUTURE WORK

Scene sketch understanding is a challenging research problem. In this work, we make the first attempt to recognize human's scene sketches. We build the first scene sketch dataset "Scene250" and open it to public. We propose, implement, and test a novel deep CNN model "Scene-Net" on scene sketch recognition. We perform fine-tuning to reduce overfitting and improve robustness of our "Scene-Net" model. The experiment results show that Scene-Net outperforms other two deep CNN models (AlexNet and Sketch-a-Net) by 7.50% and 21.25% on scene recognition respectively. Future work include collecting a large-scale scene sketches dataset, optimizing Scene-Net architecture to reduce the training time, and applying our method to sketch-based 3D model retrieval.

6. ACKNOWLEDGMENTS

This work is supported by Army Research Office grant W911NF-12-1-0057 and NSF CNS grant 1305302 to Dr. Yijuan Lu and NSF grant 1018641 to Dr. Hao Jiang.

7. REFERENCES

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

[2] C. F. Herot. Graphical input through machine recognition of sketches. In *SIGGRAPH*, 1976.

[3] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. In *arXiv:1207.0580*, 2012.

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

[5] Y. LeCun, L. Bottou, G. B. Orr, and K. R. Müller. Neural networks: Tricks of the trade. In *Efficient BackProp*, pages 9–50. Springer, 2002.

[6] T. Y. Ouyang and R. Davis. Chemink: a natural real-time recognition system for chemical drawings. In *International Conference on Intelligent User Interfaces*, 2011.

[7] P. Sousa and M. J. Fonseca. Geometric matching for clip-art drawing retrieval. *Journal of Visual Communication and Image Representation*, 20(2):71–83, 2009.

[8] I. E. Sutherland. Sketch pad a man-machine graphical communication system. In *AFIPS*, 1964.

[9] Q. Yu, Y. Yang, Y.-Z. Song, T. Xiang, and T. Hospedales. Sketch-a-net that beats humans. In *arXiv:1501.07873*, 2015.