SKETCH-BASED AERIAL IMAGE RETRIEVAL

Tianbi Jiang[1], *Gui-Song Xia*[1], *Qikai Lu*[2]

[1]*State Key Lab. LIESMARS, Wuhan University, Wuhan 430079, China*
[2]*EIS, Wuhan University, Wuhan 430079, China*
{jiangtianbi, guisong.xia, qikai_lu}@whu.edu.cn,

ABSTRACT

Notwithstanding aerial image retrieval is an important and obligatory task, existing retrieval systems lose their efficiency when there is no available aerial image used as the exemplar query. In this paper, we take free-hand sketches into consideration and address the problem of sketch-based aerial image retrieval. This is an extremely challenging task due to the complex surface structures and huge variations of resolutions of aerial images, and few works have been devoted to it. For the first time to our knowledge, we propose a framework to bridge the gap between sketches and aerial images. Specifically, an aerial sketch-image dataset is first collected. Sketches and aerial images are augmented to varied levels of details and used to train a multi-scale deep hierarchical model. The fully-connected layers of the deep model are used as cross-domain features, and the similarity between aerial images and sketches is measured by the Euclidean distance. Experiments on several public aerial image datasets demonstrate the efficiency and superiority of the proposed method.

Index Terms— Sketch, aerial image retrieval, multi-scale deep model

1. INTRODUCTION

With the development of satellite imaging sensors, the quality and amounts of aerial images increase rapidly, which brings great challenges to remote sensing image interpretation [1–3], among which it is required to browse the tremendous amounts of remote sensing images and retrieve the images we need. Recently, content-based aerial image retrieval turns to be mainstream [4, 5]. For instance, Liu *et al.* [1] built a region-based semantic feature representation to retrieve aerial images. However, since these methods all require an existing aerial image as the input, they lose efficiency when no available image samples are at hand.

Sketches are intuitive to humans and have been used to depict visual world since prehistoric times [6, 7], and sketch-based image retrieval has turned to be a practical form of retrieval. For instance, Eitz *et al.* [8] and Hu *et al.* [9] both employ modified *histogram-of-gradient* (HOG) descriptors combined with mid-level encoding to retrieve natural images. Recently, convolutional neural networks (CNNs) have also been used to accomplish sketch-based image retrieval and shown superior performance [10–12]. In a word, it has been demonstrated that cross-domain image-sketch comparison can overcome the problem of image retrieval without available query image.

Inspired by the sketch-based retrieval of natural images, this article proposes to bring free-hand sketches into the aerial image retrieval task. A key challenge in this task, however, is the complex surface structures and huge variations of image resolutions and orientations, which makes the domain gap more difficult to be bridged and the methods worked well on natural images usually fail. More precisely, we address the problem of sketch-based aerial image retrieval, where semantic sketches are used as queries to retrieve aerial scene images. We first collect an aerial sketch-image dataset, *called* Aerial-SI. Specifically, we fine-tune the Alexnet [13] using a natural sketch-image dataset to obtain a preliminary model. Sketches and images in Aerial-SI dataset are augmented to corresponding scales which posses of varied levels of details, and then used respectively to retraining the preliminary model. Thus we gain a multi-scale network. The fully-connected layers in each scale of network are connected as cross-domain features, hence we obtain the multi-scale deep model and get a domain invariance representation. Euclidean distance is employed to measure the similarity between aerial images and sketches. Experiments on public aerial scene image datasets, e.g. UCM [14] and RS19 [15] have been done to demonstrate the proposed method.

In the rest of paper, Section 2 presents the proposed framework, Section 3 explains the collection of aerial sketch-image dataset and discusses the experiment results, and Section 4 finally draws some concluded remarks.

2. METHODOLOGY

2.1. General idea

For a given query sketch S and a set of candidate aerial images $\mathcal{I} = \{I_n\}_{n=1,\dots,N}$, for sketch-based image retrieval, we

This work was in part supported by the National Natural Science Foundation of China under Grant 41501462 and Grant 91338113.

need to compute the similarity between S and $I_n \in \mathcal{I}$ to rank the images in $\mathcal{I}$ with the hope that the true match is ranked at the top. To accomplish this task, we learn a cross-domain feature and employ a good similarity measure.

2.1.1. *overview of our method*

To learn this cross-domain feature representation, inspired by [11], we preliminarily use the natural sketch-image dataset they contribute to fine-tune Alexnet. To overcome the ambiguity inherent in sketches, the collected aerial sketch-image dataset, which contains both sketches and images in a category, is processed to various levels of details. Considering the lack of data, the dataset is augmented to prevent over-fitting. Then the augmented dataset with various scales is used to retraining the preliminary model in each scale and obtain a multi-scale network. The fully-connected layers in each scale of the network are connected as cross-domain features, and the similarity between sketches and images is measure by Euclidean distance. The overview flow chart of our method is shown in Fig. 1, and the details are explained as follows.

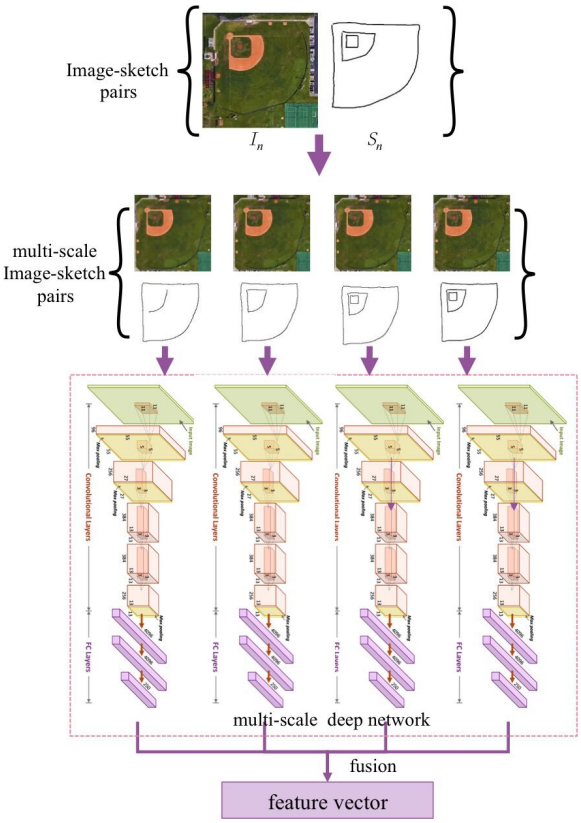


Fig. 1. Learning a multi-scale aerial image representation model. The image-sketch dataset with various levels of details are used to train a multi-scale deep network. The outputs of model are cross-domain visual features.

2.1.2. *Aerial-SI: annotated aerial sketch-image dataset*

We contribute an aerial sketch-image dataset Aerial-SI which contains 400 sketches and 3300 aerial images with 10 categories including airport, baseball field, beach, bridge, playground, pond, river, stadium, storage tanks and viaduct. Aerial images are selected from AID dataset [16, 17] and sketches are depicted by 25 volunteers who are amateur at painting. Some examples of Aerial-SI are displayed in Fig. 2. Public aerial scene image datasets UCM [14] and RS19 [15] are set as testing datasets. Specifically, UCM contains 21 land-use classes and each category has 100 aerial images, while RS19 is a dataset with 19 classes and contains 50 images for each class.

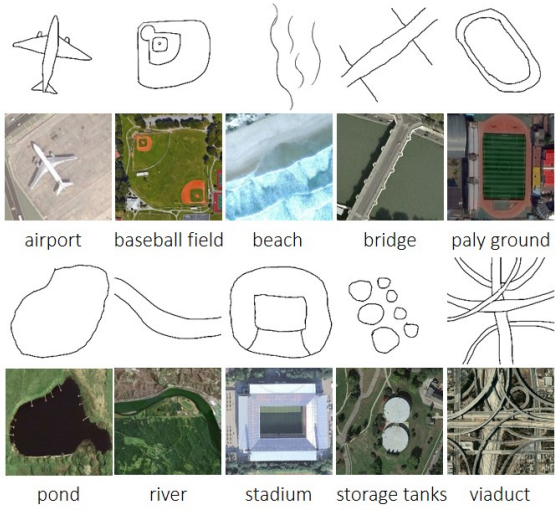


Fig. 2. Some samples of sketch-image pairs in Aerial-SI.

2.2. Learning cross-domain feature representation

Aerial images which contains complex structure and rotation brings great difficulty to retrieving aerial images using semantic sketches. Various degrees of complexity in sketches makes this task even more challenging. We aimed at learning a deep model that can extract cross-domain features and bridge the gap between aerial images and sketches.

The Alexnet [13] trained on ImageNet dataset is a powerful classifier with a strong ability to obtain image features. Thus we set Alexnet as initial value to speed up the training of our model. We use a natural sketch-image dataset, which contains both sketches and images in each category, to fine-tune Alexnet and thus obtain a preliminary model. Setting the corresponding sketch and natural image as same category help to shrink the domain gap between them. Since aerial images are different with natural images on viewing angle, structure, and various resolution, they are much more complex and are always with larger inner-class distance. We expect to overcome the problem by training a deep model on the basis of

an aerial image-sketch dataset. However, the size of aerial sketch-image dataset is too small to train a deep network. Hence, We perform the data augmentation with reflection and rotation before training. Then we add the edges of aerial images to corresponding classes to enhance the domain invariance. The processed aerial image-sketch dataset, which set corresponding images, edges and sketches as same category, is used to retrain preliminary model. We finally gain a single cross-domain deep network, and the 7th fully-connected layer of the network is set as cross-domain representation.

2.3. Multi-scale network architecture

Since the inherent ambiguity and abstraction of sketches, same category can be drawn with hugely varied levels of details. Thus we propose a multi-scale network architecture to overcome this problem.

The regular pattern for human to depict a sketch is to draw the outline strokes firstly [6], and these strokes have more descriptive power than the detailed ones. Therefore, we rank the sketch strokes according to the length, and take the top 20%, 40% , 80% and 100% to obtain sketches with 4 varied levels of details. Correspondingly, the aerial images are blurred to 4 scales using rolling guidance filter [18]. We mixing the corresponding sketches and images respectively at each scale and train the network as described in section 2.2. The 7th fully-connected layers in each scale are concatenated as output cross-domain features. Thus we obtain a multi-scale deep cross-domain model as shown in Fig. 1.

2.4. Retrieving aerial images by free-hand sketches

The original output of each scale in the network is the probability of which input aerial image belongs to each class. We remove the soft-max layer and concatenate the 7th fully-connected layers of each scale as the output of the multi-scale deep model. Hence, multi-scale cross-domain features are obtained through this deep model. The cross-domain features of query sketch and the aerial image dataset to retrieve are extracted using multi-scale deep model. In order to measure the similarity of sketches and aerial images, Euclidean distance between their features are calculated. The aerial image dataset is ranked according to the similarity and top ones are set as retrieval results. In this way we achieve the goal to retrieve aerial images by semantic free-hand sketches.

3. EXPERIMENTAL RESULTS

3.1. Experiment settings

To obtain aerial images with varied details, we set the spatial parameter σ_s of rolling guidance filter as 3, 4 and 5. Data augmentation is achieved by rotating to 9 angles in the range of $[-45^\circ$,$+45^\circ]$ degrees and reflecting. Thus the data size turns to be 18 times of the original.

During training the model, the dataset is randomly divided into 3 parts, twice of which are set as training set while the left is set as validation set. The preliminary model is trained from the 5th convolutional layer and the learning rate is set to a small value 0.001.

3.2. Baselines

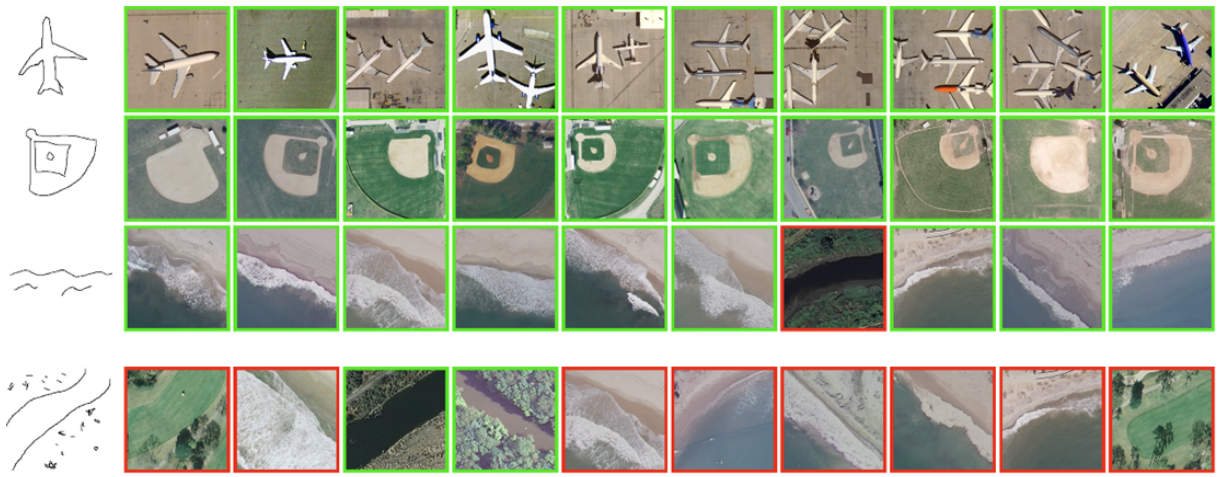
We compare our multi-scale deep model with several typical hand-crafted feature baselines including:

- **GIST:** The GIST [19] descriptor is employed and the similarity is measured by Euclidean distance.
- **BoW:** A BoW descriptor is generated with Dense-SIFT [20] to represent images and sketches. Then the descriptors are compared using histogram intersection pyramid matching kernels [21].
- **SIFT+SPM:** The 3 levels spatial pyramid [22] combined with Dense-SIFT descriptors is employed. The dictionary is set as 200 visual words and similarity is measured by histogram intersection.
- **GF-HOG:** We gain gradient field HOG [9] cooperated with BoW as feature descriptor, and codebook is set as 1000. The similarity is compared using histogram intersection.
- **DSF [11]:** HUST-SI is employed to fine-tune Alexnet and obtain a model. Euclidean distance of the 7th fully-connected layer is used to measure the similarity.

3.3. Results

We assess the proposed method on the commonly used aerial scene image datasets UCM and RS19. Since aerial scene images without salient objects, for instance, meadow and forest are improper to draw a sketch, we choose 11 classes in UCM and 6 in RS19 to depict and retrieve. For each category with salient objects, we collect 5 query sketches to evaluate the multi-scale deep model.

Retrieval results that generated from cross-domain features are presented in Fig. 3. The first column shows the query sketch and the following are corresponding top 10 results. Green box means true while the red means false. Obviously, the proposed method provide satisfactory retrieval performance on both dataset. We notice that the geometric construction of sea wave and riverbank is similar, hence the river cannot be retrieved very well. To demonstrate the proposed method, we calculate the mean average precision and top k precision. Table 1 shows the comparative results of baselines testing on UCM dataset. The last two rows are results using our single-scale deep model and multi-scale deep model. Our model performs best and the gap between our model and the baselines is large. The evaluation index proved the superiority of the proposed method.



(a) Results on UCM

(b) Results on RS19

Fig. 3. Retrieval results using multi-scale deep model: the first column shows the query sketches and each row shows top 10 results of retrieving, the green box is true while the red is false.

Table 1. Comparative results on Baselines

Features	MAP	Top-10	Top-50	Top-100
GIST	0.1305	0.1455	0.1291	0.1185
BoW	0.0517	0.0673	0.0564	0.0571
SIFT+SPM	0.0477	0.0418	0.0415	0.0402
GF-HOG	0.0752	0.1200	0.1022	0.0855
DSF	0.3038	0.5418	0.3956	0.3125
OURS(SINGLE)	0.4455	0.7073	0.5487	0.4407
OURS(MULTI)	**0.4636**	**0.7218**	**0.5673**	**0.4582**

From the results, we find that sketches which seem like abstract of image can represent contextual and structural information of aerial scene. The proposed model which is trained using mixing data can obtain the common feature of sketches and aerial images, thus it shrinks the distance between them in feature space and obtains a superiority result.

4. CONCLUSION

In this paper, we proposed a multi-scale deep model for sketch-based aerial image retrieval task. An aerial sketch-image dataset Aerial-SI with annotation is contributed to stimulate the research. Achieving the task requires a multi-scale deep model learned with augmented aerial sketch-image dataset. The output of multi-scale model is cross-domain features, and similarity are measured by Euclidean distance between features of query sketch and aerial images. Experiments on commonly used aerial scene image datasets confirm the efficiency and superiority of proposed method, and demonstrate that our model can bridge the domain gap between sketches and aerial images.

5. REFERENCES

[1] Tingting Liu, Liangpei Zhang, Pingxiang Li, and Hui Lin, "Remotely sensed image retrieval based on region-level semantic mining," *EURASIP Journal on Image and Video Processing*, vol. 2012, no. 1, pp. 4, 2012.

[2] Fan Hu, Gui Song Xia, Zifeng Wang, Xin Huang, Liangpei Zhang, and Hong Sun, "Unsupervised feature learning via spectral clustering of multidimensional patches for remotely sensed scene classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 8, no. 5, pp. 2015–2030, 2015.

[3] Fan Hu, Gui-Song Xia, Jingwen Hu, and Liangpei Zhang, "Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery," *Remote Sensing*, vol. 7, no. 11, pp. 14680–14707, 2015.

[4] Zhongxiang Du, Xuelong Li, and Xiaoqiang Lu, "Local structure learning in high resolution remote sensing image retrieval," *Neurocomputing*, 2016.

[5] Gui-Song Xia, Julie Delon, and Yann Gousseau, "Shape-based invariant texture indexing," *International Journal of Computer Vision*, vol. 88, no. 3, pp. 382–403, 2010.

[6] Mathias Eitz, James Hays, and Marc Alexa, "How do humans sketch objects?," *Acm Transactions on Graphics*, vol. 31, no. 4, pp. 44, 2012.

[7] Xiang Bai, Song Bai, Zhuotun Zhu, and Longin Jan Latecki, "3d shape matching via two layer coding," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 12, pp. 2361–2373, 2015.

[8] M Eitz, K Hildebrand, T Boubekeur, and M Alexa, "Sketch-based image retrieval: Benchmark and bag-of-features descriptors.," *IEEE Transactions on Visualization Computer Graphics*, vol. 17, no. 11, pp. 1624–1636, 2011.

[9] Rui Hu and John Collomosse, "A performance evaluation of gradient field hog descriptor for sketch based image retrieval," *Computer Vision Image Understanding*, vol. 117, no. 7, pp. 790–806, 2013.

[10] Qian Yu, Feng Liu, Yi-Zhe Song, Tao Xiang, Timothy M. Hospedales, and Chen Change Loy, "Sketch me that shoe," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[11] Xinggang Wang, Xiong Duan, and Xiang Bai, "Deep sketch feature for cross-domain image retrieval," *Neurocomputing*, 2016.

[12] Wei Shen, Xinggang Wang, Yan Wang, Xiang Bai, and Zhijiang Zhang, "Deepcontour: A deep convolutional feature learned by positive-sharing loss for contour detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3982–3991.

[13] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 25, no. 2, pp. 2012, 2012.

[14] Yi Yang and Shawn Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*. ACM, 2010, pp. 270–279.

[15] Gui-Song Xia, Wen Yang, Julie Delon, Yann Gousseau, Hong Sun, and Henri Maître, "Structural high-resolution satellite image indexing," in *ISPRS TC VII Symposium-100 Years ISPRS*, 2010, vol. 38, pp. 298–303.

[16] Jingwen Hu, Tianbi Jiang, Xinyi Tong, Gui-Song Xia, and Liangpei Zhang, "A benchmark for scene classification of high spatial resolution remote sensing imagery," in *2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. IEEE, 2015, pp. 5003–5006.

[17] Gui-Song Xia, Jingwen Hu, Fan Hu, Baoguang Shi, Xiang Bai, Yanfei Zhong, and Liangpei Zhang, "Aid: A benchmark dataset for performance evaluation of aerial scene classification," 2016.

[18] Qi Zhang, Xiaoyong Shen, Li Xu, and Jiaya Jia, "Rolling guidance filter," in *European Conference on Computer Vision*. Springer, 2014, pp. 815–830.

[19] A Oliva and A Torralba, "Building the gist of a scene: the role of global image features in recognition.," *Progress in Brain Research*, vol. 155, no. 2, pp. 23–36, 2006.

[20] David G. Lowe, "Distinctive image features from scale-invariant keypoints," in *International Journal of Computer Vision*, 2004, pp. 91–110.

[21] Kristen Grauman and Trevor Darrell, "The pyramid match kernel: Discriminative classification with sets of image features," in *Tenth IEEE International Conference on Computer Vision*, 2005, pp. 1458–1465 Vol. 2.

[22] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce, "Spatial pyramid matching," *Cambridge University Press*, 2009.