DEIM Forum 2019 F6-2

Discriminative Objects Discovery for Spot Assessment

陳　天偉† 馬　強†

† School of Informatics, Kyoto University
Yoshida-honmachi, Sakyo-ku, Kyoto, 606–8232, Japan
E-mail: †chentianwei@db.soc.i.kyoto-u.ac.jp, ††qiang@i.kyoto-u.ac.jp

Abstract Discovering discriminative objects that has local characteristics is an important task for sightseeing spot discovery and assessment. In previous work, discriminative objects are detected in the early stage of whole process and thus can hardly get discriminative objects when target spots change. Such inflexibility makes it hard to discover reasonable discriminative objects and will cause a large computational cost for further process. In this paper, we propose a new framework of discriminative objects discovery. It detects common objects from each sightseeing spot, ang then generates hierarchical clusters of these objects. Our framework leaves the process of discovering discriminative objects at the last stage when we want to compare several target spots. When a new requirement of discover discriminative objects comes, the only thing the proposed framework should do is to compare data from two tree-like clusters and find what the discriminative objects are.

Key words Sightseeing Spot Assessment, Discriminative Object Discovery, Unsupervised Object Detection

1 Introduction

Beautiful sceneries, amazing events and memorable experience, travel is such an attractive activity that draws people ' s attention and comes to an important part of people ' s daily life. However, there are always the problems like which spots are worth to go and why they are of worth, which confuse people who want to have a travel.

Sightseeing spot assessment is the research topic focusing on this problem, which is to give an assessment to sightseeing spots. Current methods of sightseeing spot assessment [1, 2, 7, 8, 18, 21] are focusing on assessment image data from spots, which can reflect sightseeing spot ' s quality directly. However, the utilization of images shared by users is a challenging task, as many factors like the quality of the image and the relevance between images and spots could affect the result deeply.

One reasonable solution of utilizing image data to assessment sightseeing spot is to discover discriminative elements (e.g. patches [7] or objects [8]) from those images. The core idea is that by comparing images from different spots, we can discover some elements that appears much in a specified spot and do not appear in others. We thus called them discriminative elements. By observation, those elements are of high local characteristics, which reveal the spot ' s uniqueness and could be a clue to evaluate the worth of sightseeing spot. Also, since the process only concerns about some valuable subpart of images and remove common elements by comparing target spots, this idea has high robustness against noise and can get much more reasonable outputs.

But since the processing of finding discriminative elements of the previous methods is conducted in the early stage, the whole framework become inflexible to adapt new data. E.g. Those discriminative objects can hardly represent the difference between Kyoto and Nara, but if we add Nara as a candidate city to find discriminative objects, we have to run the whole process again. It is not the thing we expect, as it is impossible to collect all sightseeing spots around the world and send them in one training. In short, it is of time-consuming and laborious to retrain such a complex model when new data come.

Also, we consider that the discriminability in object-level cannot represent a spot ' s characteristics and internal universality well. E.g. the building in Kyoto and Nara is very similar in the shape, which makes them hard to distinguish in object-level. However, if we compare them more carefully, we may find that most of them are in different style, which can distinguish them more precisely. Thus, a tinier scale should be considered to represent the discriminability of each spot.

In this paper, we propose a new framework of discriminative objects discovery from sightseeing spots to solve the problem above. It discovers objects from geo-tagged images in an unsupervised manner and avoids finding discriminative objects in the early steps. To represent a spot ' s characteristics well, we divide the objects into several patches, which can represent more discriminative details and is more human-interpretable. We sampling these millions of patches

and try to reserve those key ones as much as possible. Furthermore, without the limitation of extract discriminative elements from candidates in the first step, we can collect and organize more data for further use.

2 Related works

Our work is mainly about sightseeing spot assessment and based on the idea of extract discriminative elements from images by utilizing unsupervised object discovery method. Thus, we would introduce those related works before our proposed method.

2 1 Object detection

Object detection is a fundamental and challenging research topic which may bring benefits to several related topics and applications. In recent year, object detection methods based on CNN made a great breakthrough in this research task [9, 10, 11]. Those methods utilized large amount of labeled image data to train a CNN based model in supervised manner, which make the model have the capacity to proposed regional bounding box to new data. However, due to the supervised training manner, the model usually can only detect objects that given in the training set. As a matter of fact, in many research and application scenario, the data with labeled region and object are not so easy to get, which make them hard to share the benefits from recent breakthrough. There are also some research topics focusing on this problem and try to propose solutions, such as Weak Supervised Object Detection [12, 13] and Zero-shot Object Detection [14, 15].

However, the methods mentioned above still aims to detect limited categories of objects and more or less rely on human annotation.

As our task aims to discover unlimited categories of object and only consider if the object is discriminative, we prefer unsupervised object detection methods to get objects we need. In this research topic, Tang et al. [16] proposed a joint image-box formulation to discover unlabeled objects. Cho et al. [17] apply a probabilistic Hough matching method to deal with the task of discover and localize objects.

2 2 Sightseeing spot assessment

Sightseeing spot assessment is the task to give a reasonable evaluation score to sightseeing spot. This task is highly related to the research of POI (place of interest) discovery [3, 4] and recommendation [3, 4, 5] as the assessment could be a kind of clue to discover POIs. However, sightseeing spot assessment only focuses on evaluate sightseeing spots while POI discovery not only considers about sightseeing spot, but also hotels, restaurants and other spots.

Currently, many methods [3, 8, 18] are focusing on estimate sightseeing spots by geo-tagged images directly, which avoid affects from the gap between popularity and sightseeing quality, and can apply to estimate obscure spots.

2 3 Discriminative sightseeing elements discovery

Discovering discriminative sightseeing elements is a subtopic of tourism analysis, which aims to discover elements in one spot that make the spot distinguishable to other spots.

In this topic, Doersch et al.[7] proposed a method to discover discriminative patches of Paris from Google Street View. By iteratively comparing Paris image data and other cities image data, their model output several patches which appear in Paris and do not appear in other candidate cities, thus called the things “make Paris look like Paris”.

Following this idea of discovering discriminative elements, Ge et al.[8] proposed a robust visual object clustering method and apply it to sightseeing spot assessment. In this paper, they use the manner of iteratively comparing images from Kyoto and other four cities, then return the discriminative objects of Kyoto. The result of discriminative objects include temples, towers, traditional houses and maple trees, which depict characteristics of sightseeing spots in Kyoto.

These methods have shown a good performance. In this research topic, Ge et al. [8] proposed a discriminative objects discovery framework to discover local characteristics in Kyoto by iteratively comparing Kyoto and other cities. Although the data has lot of noise, the result still clearly shows a lot of Kyoto’s local characteristics.

3 Methodology

There are two core ideas in our work. One idea is to avoid discovering discriminative objects in the early step, which makes the framework more flexible when target spots changing. Another idea is to divide objects into patches and discover discriminative patches to represent objects. This method can reserve more details of one spot than use object directly, and can calculate the discriminability in a more reasonable and human-interpretable way.

Based on these two ideas, we propose a new framework which use an unsupervised method to discover candidate objects for each spot, divide objects to patches, extract features and use clustering method to find the common elements from patches to represent objects and spots. When the requirement of find discriminative objects comes, we then start to compare target spots to find the discriminative objects. The framework and data flow are showed in Figure 1.

There are mainly two parts in our framework, one is mianly about preparing common elements (objects and patches), another one is to discover discriminative elements. In the first part, images from each city are processed into patches that are common in that city. In the second part, these patches will be winnowed down to a set of patches that with high discriminability to a city. Then, the discriminative patches

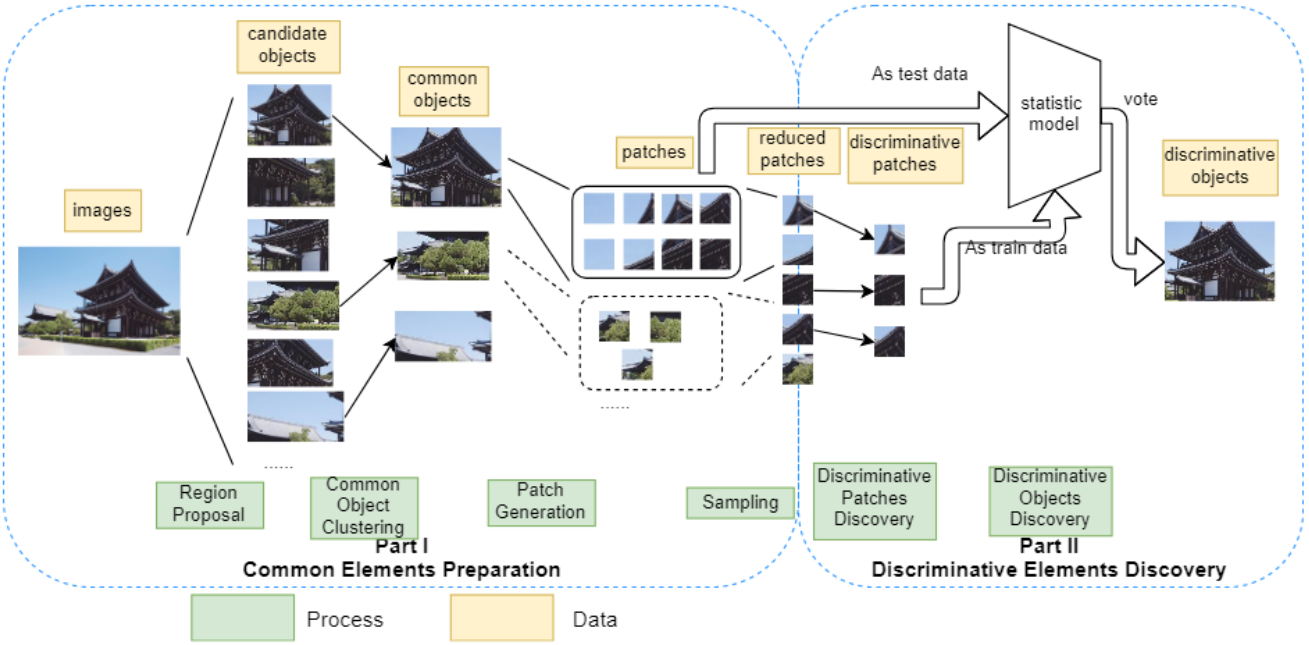

Figure 1 Framework & data flow. The brown box represent what kind of data they are, and the green box represent the processes we do.

are used to train a model to find the most discriminative objects in one city.

To be noticed that as we don't extract discriminative objects in the first step, the process logic changes and each component are defined with different methods. Thus, the whole framework is different from previous works.

3.1 Region Proposal & Common Object Clustering

According to our task definition, the discriminative objects can be everything that make the candidate distinguishable, i.e. we cannot apply any prior knowledges such as which categories of object might be discriminative on while which might not. Thus, we first apply a well-known unsupervised region proposal method [19] to propose every possible object from geo-tagged images.

To improve the quality of proposed region and for further use, we only allow the region that larger than $[64 \times 64]$ but smaller than 90% pixels of the whole image to be valid and only select the largest one if several regions overlap to each other. After the process, we simply treat those regions as candidate objects.

By observation, there are a lot of candidates don't catch object, or only catch a part of an object. Thus, we use Kmeans algorithm to cluster those candidate objects. More detailed, we use a pre-trained VGG-19 [20] model to extract features from each patch. In the extraction process, we convert all the candidates into $[64 \times 64]$ shape and get features from VGG-19's last pooling layer, which still retain the location information for each object and thus make the clustering more reasonable. After that, we cluster all candidates in one image to limite that each image can only output 3 candidate objects, which is from the top-3 largest clusters. By this process, we make the output candidate objects more clear. Then, we use Kmeans again to cluster the candidates which may still depict same object while from different images. From each cluster, we select only one as a representative of this cluster. The volume of candidate object thus reduce slightly. After the process above, we treat these objects as common objects which appear requently in one city.

3.2 Patch Generation & Sampling

By the view that the discriminability in object-level directly cannot represent a spot's characteristics and internal universality well, we divide the objects into several patches to reserve more details for later processing. Thus, we use a sliding window with the size of $[32 \times 32]$, and set the strid to 8, to divide each object into several patches. To be noticed that we orgatized the patches by object index to reserve the belonging information for discovering discriminative objects, and thus in Figure 1, the patches from one object are blocked in one box.

In this process, millions of patches are generated from common objects and make it almost impossible to do any process later. Thus, we randomly sample a reasonable number of patches to represent a city. In this paper, we sample about 2000 patches for each city.

Here, we use a pre-trained VGG-19 [20] model to extract features from each patch. The size of VGG-19's output is 512, we treat each of them as a vector which represent a

patch.

3 3 Discriminative Patch Discovery

With the clusters generated in last process, get a lot of patches which come from common objects. However, those patches can only represent the things appears frequently in one city, but are not discriminative, as they may also appears a lot in other city. Thus, we try to compare them with patches from other cities and to find those discriminative patches from them. The whole process are showed in Algorithm 1.

Algorithm 1

Require:

Positive set P, Negative set N;

Begin:

1 : $P \rightarrow \{P1,\ P2\}$;

2 : $P_{train} \leftarrow P1$; $P_{next} \leftarrow P2$;

3 : *While not converged*() *do*

4 : $Model \leftarrow SVM_{train}([P_{train},\ N])$;

5 : $P_{new} \leftarrow select(Model,\ P_{next},\ > \gamma)$;

6 : $swap(P_{new},\ P_{next})$; $init(Model)$;

7 : *Endwhile*

8 : *Return* $[P_{new}, P_{next}]$

There are two steps in this process. At first, we treat the patches from target city as positive set P, while those patches from other cities are treated as negative set N. Then, we divide the positive set into two equal, non-overlapping subset P_1 and P_2 for cross-validation.

Given a positive subset P_1, we concatenate P_1 with negative set N as training set, and the data from positive set are assigned 1 and those negatives are assigned -1. We then initialize and train a SVM model on this training set. When the training finished, we send other positive set P_2 as test set and let the model to give the probability of the object belongs to positive set. Here we set a threshold parameter γ . If an object ' s probability is smaller than γ , we treat the object is not enough discriminative and discard it in later processing.

After the process, we get a purer positive set P_2'. Then, we swap P_1 and P_2', concatenate P_2' with negative set N as a new training set and use it to winnow P_1. This process goes iteratively until convergence, i.e. the total number of positive objects doesn ' t change anymore. At the time iterative process ends, we can get the discriminative patches.

3 4 Discriminative Objects Discovery

In last process, we get a large amount of patches which are discriminative and all comes from one city:s object. However, either retrieve objects by patches or combine patches to objects is almost impossible, as we only randomly sample a few of the patches from millions of them. Here, our idea to discover discriminative objects is to train a new SVM model as a detector of object's discriminability.

More concretely, we set those discriminative patches from one city as positive set, and set those discriminative patches belong to other cities as negative set. Then, we initial a new SVM model and use the data above to train it. After training, we load patches which are generated in section **3 2** as test data of the SVM model. Those patches are organised by object index and thus we can easily find which object the patch belongs to.

In this processing, we treat those patches with the posibility higher than 50% as discriminative patches and calculate the radio of those patches in one object to detect discriminative object. In this paper, if more than half patches in one object are discriminative, we then treat this object as a discriminative one. The calculation of discriminative radio in object O_n is as follow:

$$O_n = P_p / P_all$$

Where P_p represents the number of discriminative patches and P_{all} represents the number of patches belong to object O_n.

4 Experiment

4 1 Dataset

To verify the utility of our method and compare the performance with previous work, we first use the dataset provided by [8], which consisted of photos of sightseeing spots from five cities: Kyoto, Xi ' an, Beijing, Paris and San Francisco. The dataset collected from approximate 30 most popular sightseeing spots from TripAdvisor for each city and 50-100 images are downloaded per spot from Flickr's API. The final dataset for each city are formed by randomly select 1000 image from downloaded images.

Additionally, for showing our framework ' s flexibility, we collect data from a new city, Nara, in the manner of original dataset did.

4 2 Result & Evaluation

To evaluate the performance of our framework, we are focusing on three main part: the usability, the accuracy and the flexibility. These three part can derive to three questions:

1). Does the model find the discriminative objects?

2). Can we quantitively show the accuracy?

3). Does the model flexibility?

In the later parts of this section, we do the evaluation based on these three question.

4 2. 1 The disxriminability of objects

During this part, to make an evaluation of discriminability of objects discovered by our framwork, we randomly output 10 discriminative objects of them, which have a discriminative radio higher than 0.95. Figure 2 shows the 10 discriminative objects from Kyoto and the other cities are Beijing, Paris, San Francisco, Xi'an and Nara.

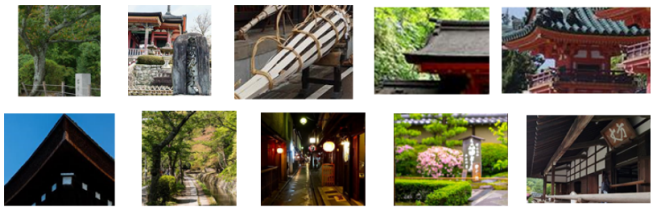

Figure 2 10 high discriminative radio objects from Kyoto

The result shows some objects that we can usually see in Kyoto, and most of them are discriminable if comparing with images from other cities.

4 2. 2 The qualify of discriminability

As giving quantitive evaluation to discrimiability is a difficult, we set the experiment as to test a new unseen dataset with 2000 patches, which are 50% from target city and 50% from other cities.

To make a comparation of Doersch et al.[7] , we implement their key component of discovering discriminative patches in Python and run our collected dataset throw the component. Since Ge et al.[8] simpliy apply the model in Doersch et al.[7], we treat that they use the same component to discover the discriminative patches and only compare the AUC result with Doersch et al.[7]'s work. The result are as follow:

Table 1 Result of evaluating patches' discriminability (AUC)

City	Ours	Doersch's
Kyoto	0.772	0.816
Xi ' an	0.719	0.798
Beijing	0.662	0.647
Paris	0.854	0.954
San Francisco	0.697	0.696
Nara	0.753	0.798

4 2. 3 The analysis of flexibility

In Ge et al.[8]'s work, the process of discovering discriminative objects starts as the first step, which makes the whole framework and its intermedia outputs can hardly be used for comparing other candidates. E.g. If a tourist from Beijing comes Kyoto, the framework will process the pair {Kyoto, Beijing} and discovers the discriminative objects that appears a lot in Kyoto while not in Beijing. While a tourist come from Paris then comes to Kyoto, the framework have to run again to find discriminative objects by the pair {Kyoto, Paris}. And even a person who has been to Beijing and Paris (as the pair {Kyoto, [Beijing, Paris]}) comes, the whole framework have to run again to discover discriminative objects.

Assume there will be S groups of tourist with different background (experiments of other cities) come to Kyoto and K spots are overlapped. Then, we have to run the whole framework S times to get all result.

In our framework, we first make the process of finding common objects and patches save them for later process. When a query (e.g. pair {Kyoto, Beijing}, {Kyoto, Paris} or {Kyoto, [Beijing, Paris]}) comes, we then find the discriminative objects. Assume we face the situation the same as above, our framework only need to do the time consuming process of preparing common objects and patches $S - K$ times. Thus, our model are more flexible than Ge et al.[8]'s.

5 Conclusion

In this paper, we propose a new discriminative object discovery framework which solve two problem of previous works. The first one is about the inflexibility of previous works due to the discriminative element discovery processing. The second one is during the process, we try several methods to reduce the cost while avoid discovering discriminative objects in the early step.

However, current discriminative objects are still hard to show a good performance, as some of them may also appears a lot in other cities (e.g. pine tree also appears a lot in Beijing and Nara). The problem may caused by the volume of the data we collect and the sampling strateegies we used, which may lose some key discriminative information before we train the model. Thus, in the later work, we may collect more data from each spot and try other sampling strategies that may reserve more discriminative information.

This work is partly supported by MIC SCOPE(172307001).

References

[1] Chen, Wei-Chao, Agathe Battestini, Natasha Gelfand and Vidya Setlur, "Visual summaries of popular landmarks from community photo collections," Conference Record of the Forty-Third Asilomar Conference on Signals, Systems and Computers, 1248-1255, 2009.

[2] Berman, Marc G., Michael C. Hout, Elizabeth A Necka, MaryCarol R. Hunter, Grigori Yourganov, John M. Henderson, Taylor Hanayik, Hossein Karimi and John Jonides, "The Perception of Naturalness Correlates with Low-Level Visual Features of Environmental Scenes," PloS one, 2014.

[3] Hasegawa, Keisuke, Qiang Ma and Masatoshi Yoshikawa,

"Trip Tweets Search by Considering Spatio-temporal Continuity of User Behavior," DEXA, 2012.

[4] Zhuang, Chenyi, Qiang Ma, Xuefeng Liang and Masatoshi Yoshikawa, "Anaba: An obscure sightseeing spots discovering system," 2014 IEEE International Conference on Multimedia and Expo (ICME), 1-6, 2014.

[5] Zhao, Pengpeng, Xiefeng Xu, Yanchi Liu, Victor S. Sheng, Kai Zheng and Hui Xiong, "Photo2Trip: Exploiting Visual Contents in Geo-tagged Photos for Personalized Tour Recommendation," ACM Multimedia, 2017.

[6] Lim, Kwan Hui, Jeffrey Chan, Christopher Leckie and Shanika Karunasekera, "Personalized Tour Recommendation Based on User Interests and Points of Interest Visit Durations," IJCAI, 2015.

[7] Doersch, Carl, Saurabh Singh, Abhinav Gupta, Josef Sivic and Alexei A, "What makes Paris look like Paris?," ACM Trans. Graph. 31, 101:1-101:9, 2012

[8] Min Ge, Chenyi Zhuang, Qiang Ma, "A Ranking Based Approach for Robust Object Discovery from Images of Mixed Classes," Asia Information Retrieval Societies (AIRS), 71-83, 2017.

[9] Girshick, Ross B., Jeff Donahue, Trevor Darrell and Jitendra Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," 2014 IEEE Conference on Computer Vision and Pattern Recognition, 580-587, 2014.

[10] Girshick, Ross B., "Fast R-CNN," 2015 IEEE International Conference on Computer Vision (ICCV), 1440-1448, 2015.

[11] Ren, Shaoqing, Kaiming He, Ross B. Girshick and Jian Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," IEEE Transactions on Pattern Analysis and Machine Intelligence 39, 1137-1149, 2015.

[12] Dong, Xuanyi, Liang Zheng, Fan Ma, Yi Yang and Deyu Meng, "Few-Example Object Detection with Model Communication," IEEE transactions on pattern analysis and machine intelligence, 2018.

[13] Teh, Eu Wern, Mrigank Rochan and Yang Wang, "Attention Networks for Weakly Supervised Object Localization," BMVC, 2016.

[14] Bansal, Ankan, Karan Sikka, Gaurav Sharma, Rama Chellappa and Ajay Divakaran, "Zero-Shot Object Detection," ECCV, 2018.

[15] Rahman, Shafin, Salman Hameed Khan and Fatih Murat Porikli, "Zero-Shot Object Detection: Learning to Simultaneously Recognize and Localize Novel Concepts," CoRR abs/1803.06049, 2018.

[16] Tang, Kevin D., Armand Joulin, Li-Jia Li and Li Fei-Fei, "Co-localization in Real-World Images," 2014 IEEE Conference on Computer Vision and Pattern Recognition, 1464-1471, 2014.

[17] Cho, Minsu, Suha Kwak, Cordelia Schmid and Jean Ponce, "Unsupervised object discovery and localization in the wild: Part-based matching with bottom-up region proposals," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1201-1210, 2015.

[18] Shen, Yizhu, Min Ge, Chenyi Zhuang and Qiang Ma, "Sightseeing Value Estimation by Analyzing Geosocial Images," 2016 IEEE Second International Conference on Multimedia Big Data (BigMM), 117-124, 2016.

[19] Uijlings, Jasper R. R., Koen E. A. van de Sande, Theo Gevers and Arnold W. M. Smeulders, "Selective Search for Object Recognition," International Journal of Computer Vision 104, 154-171, 2013.

[20] Simonyan, Karen and Andrew Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," CoRR abs/1409.1556, 2014.

[21] Shen, Yizhu, Chenyi Zhuang and Qiang Ma, "Element-Oriented Method of Assessing Landscape of Sightseeing Spots by Using Social Images," APWeb/WAIM, 2017.