Can Off-The-Shelf Object Detectors Be Used to Extract Geographic Information From Geo-referenced Social Multimedia?

Daniel Leung
University of California, Merced
5200 North Lake Rd.
Merced, CA 95343
cleung3@ucmerced.edu

Shawn Newsam
University of California, Merced
5200 North Lake Rd.
Merced, CA 95343
snewsam@ucmerced.edu

ABSTRACT

On-line photo sharing websites such as Flickr not only allow users to share their precious memories with others, they also act as a repository of all kinds of information carried by their photos and tags. The objective of this work is to perform geographic knowledge discovery by crowdsourcing of geographic information from Flickr's geo-referenced photo collections. In particular, we explore the idea of extracting geographic information semantically for land-use classification by applying state-of-the art object and concept detectors directly to the photo collections. Our results suggest that even though the detectors are able to produce distinctive spatial distributions of different objects, performing land-use classification using user contributed geo-referenced photos remains a challenging problem due to the wide variety of photos available in the collections.

Categories and Subject Descriptors

I.4.8 [**Image Processing and Computer Vision**]: Scene Analysis; I.5.4 [**Pattern Recognition**]: Applications; H.2.8 [**Database Management**]: Database Applications—*spatial databases and GIS*

General Terms

Algorithms, Experimentation

Keywords

Geographic discovery, geo-tagged, land-use classification

1. INTRODUCTION

On-line photo sharing websites such as Flickr [1] and Picasa [2] have become popular channels for people to share their precious memories with one another. Although these photo collections capture many memories, they also contain other information that may be interesting particularly in different contexts. We usually think of the 5 W's and 1 H (Who, What, Where, When, Why, and How) when we read literatures, but each of the photos in the collections can also provide us with some of these six types of information. Therefore, we can say that these online photo sharing websites act as a repository of all kinds of information. This allows individuals to perform knowledge discovery by crowdsourcing of information through these photo collections. With more than 180 million geo-referenced photos available from Flickr, our goal in this work is to map what-is-where on the surface of the Earth using the What and Where aspects of the information. In particular, we explore the idea of extracting geographic information semantically for land-use classification by applying object detectors directly to the photo collections.

The novel contribution of this work is to use proximate sensing to compliment the shortcoming of remote sensing in land-use classification. We propose a novel framework of using state-of-the-art object detectors to perform geographic discovery in large collections of geo-referenced photos. This framework can be applied to any land-use classes, especially classes that cannot be discerned by using overhead imagery such as trade, services, cultural, entertainment, and recreational facilities that usually belong to the same developed land-cover class. Sections 2 and 3 give a brief background of proximate sensing and object detection. In Section 4 we describe the dataset and experiments. The experimental results are presented in Section 5 and followed by a discussion on challenges and the conclusion in Sections 6 and 7.

2. PROXIMATE SENSING

In traditional remote sensing, overhead imagery is used to distinguish different types of land-cover in a given region; however, it has difficulty in telling the type of land-use a certain land-cover class belongs to. For example it is easy to locate a region with large buildings and parking lots in the satellite view mode in Google Maps, but it is much more challenging to use the satellite view to determine whether the region belongs to a shopping center or a warehouse. To find out the answer, one can switch to the street view mode and see the images of nearby objects and scenes taken from the ground level. While there has been some work by others on knowledge discovery from ground-level images, such as methods for discovering spatially varying (visual) cultural differences among concepts such as "wedding cake" [10] and for discovering interesting properties about popular cities

ACM SIGSPATIAL LBSN '12, November 6, 2012. Redondo Beach, CA, USA

and landmarks such as the most photographed locations [3], we use the term "Proximate Sensing" to describe a more comprehensive framework that uses ground level images of nearby objects and scenes to automatically map what-is-where on the surface of the earth similar to how remote sensing uses overhead images.

3. OBJECT DETECTION FOR LAND-USE CLASSIFICATION

In computer vision, land-use classification can be considered as a problem of image understanding. There are two commonly used approaches in solving this type of problems, low-level and high-level analysis. In low-level analysis, images are interpreted in a bottom-up direction where features are derived at the pixel level. These features such as colors, texture, and other transformations of the pixel values are used to characterize images in a statistical way. In our previous work [6, 7], we are able to demonstrate how simple low-level features from geo-referenced photos can be used to perform land-cover classification.

Although low-level analysis has been the main approach to the image understanding problems, these low-level features do not characterize the image at a semantic level. As we have mentioned the 5 W's and 1 H at the beginning, it is very difficult to extract these types of semantic information by using the pixel values of the photos. As a result, a high-level or top-down approach to this type of problem has been proposed. This approach analyses images at the level of objects, concepts, events, and activities by using different kind of detectors. There has been much progress in computer vision on object detection over the last decade. This is in large part a result of image analysis based on local invariant features which, besides the invariance properties, are robust to occlusion, a major challenge in object detection. Providing an overview of state-of-the-art techniques in object detection is beyond the scope of this paper; however, a good survey can be found in [9].

4. EXPERIMENT

Our focus in this work is to investigate whether object detectors can extract geographic information that is useful for land-use classification from the geo-referenced photo collections. As a first step, we explore whether the object detectors can produce maps of objects with distinctive spatial distributions within a study region.

Our study region is the 10x11km center of metropolitan London, UK. This region includes commercial, residential, as well as recreational areas. We divide the study region into 110 1x1km sub-regions (tiles) and collect photos according to the coordinates of each tile using the Flickr API. We then apply detectors of 177 objects to these photos. Examples of objects used are listed in Table 1.

Table 1: Object Examples

Airplane	Baseball	Candle	Duck
Ferris wheel	Flower	Goggles	Gravel
Helmet	Keyboard	Loudspeaker	Microwave
Newspaper	Pot	Roller coaster	Shield
Skyscraper	Telephone	Umbrella	Window

The object detectors we apply in this work are the Object Bank representation developed by Li et al. [8]. It is an implementation of the latent SVM detectors [4] and texture classifiers [5] for 177 objects in different scales and spatial pyramid levels. To detect an object of different sizes, we set the scale level to the maximum of 12 and select the highest detection rate value among the 12 levels as the detection rate for each object. Since our focus is to detect whether an object appears in a photo or not, the spatial location of that object is not as relevant and therefore we only consider the first level of the spatial pyramid. As a result, each photo will be represented by a distribution of detection rates of the 177 objects. A threshold value is selected for each of the objects so that a particular object is considered as present in a photo if the detection rate of this object is higher than the corresponding threshold value. To generate a map of an object, we simply count the number of photos labelled as containing the object within each geographic tile and normalize the counts by the total number of photos within the tile. This forms a distribution of that object across the tiles, hence the object map. Figure 1 shows the framework of producing object maps.

In order for the results from the object detectors to be geographically informative, maps of the detected objects should display distinctive spatial distributions. To study this behaviour, we perform co-occurrence analysis on each object map. We treat each object map as a grayscale image and evaluate its co-occurrence matrix by measuring the distribution of spatially co-occurring object counts across the study region. We then calculate the homogeneity of the co-occurrence matrix of each object. Homogeneity is a measurement of closeness of distribution of the object counts in an object map. It ranges from 0 to 1, where a 1 indicates that locations with similar number of objects detected are clustered together. Objects with less homogeneity (or more heterogeneity) suggest that these objects are not present evenly across the study region.

Besides the distinctiveness of the object distributions, it is interesting to investigate the spatial correlations between objects since related objects should appear in the same land-use region. To measure the correlation between objects we compute the correlation coefficients between the 10 objects that are the most heterogeneously distributed in the study region. Correlation coefficient ranges from -1 to 1, where a 1 (or -1) suggests that there is positive (or negative) linear relationship between the objects.

5. RESULTS

The 10 most heterogeneously distributed objects are listed in Table 2, and their corresponding object maps are shown in Figure 3. From Figure 3, we can see that these 10 objects have different spatial distributions across the study region and we believe that these spatially distinctive distributions might provide meaningful geographic information that could be useful for land-use classification.

Table 3 shows the correlation coefficients for pairs of the 10 most heterogeneously distributed objects. While we find pairs of objects such as desks and desktop computers, plates and fruits, that are related logically, we also find some illogical pairs such as clams and gallery, and plates and basketball hoops. As we further investigate this problem, we discover that the detectors are often not detecting what they are designed to detect. In other words, the false positive rate of the detectors is high. Figure 2 illustrates some of the false

Table 3: Correlation coefficients for pairs of the 10 most heterogeneously distributed objects.

	Light	Sky	Fence	Desk	Gallery	Soil	Basketball hoop	Clock	Desktop computer	Boot
Light	1.0000	0.2285	0.4196	0.1192	0.2191	0.5729	0.2406	0.2546	0.2438	0.5046
Sky	0.2285	1.0000	0.3861	0.2836	0.0650	0.3893	0.1466	0.1627	0.1838	0.3662
Fence	0.4196	0.3861	1.0000	0.4229	0.5803	0.5456	0.4455	0.4083	0.4369	0.5152
Desk	0.1192	0.2836	0.4229	1.0000	0.4428	0.1274	0.4329	0.3712	0.4853	0.5455
Gallery	0.2191	0.0650	0.5803	0.4428	1.0000	0.1691	0.4537	0.4232	0.6080	0.2406
Soil	0.5729	0.3893	0.5456	0.1274	0.1691	1.0000	0.2323	0.2108	0.1468	0.3750
Basketball hoop	0.2406	0.1466	0.4455	0.4329	0.4537	0.2323	1.0000	0.8988	0.6825	0.4016
Clock	0.2546	0.1627	0.4083	0.3712	0.4232	0.2108	0.8988	1.0000	0.5812	0.3565
Desktop computer	0.2438	0.1838	0.4369	0.4853	0.6080	0.1468	0.6825	0.5812	1.0000	0.5347
Boot	0.5046	0.3662	0.5152	0.5455	0.2406	0.3750	0.4016	0.3565	0.5347	1.0000

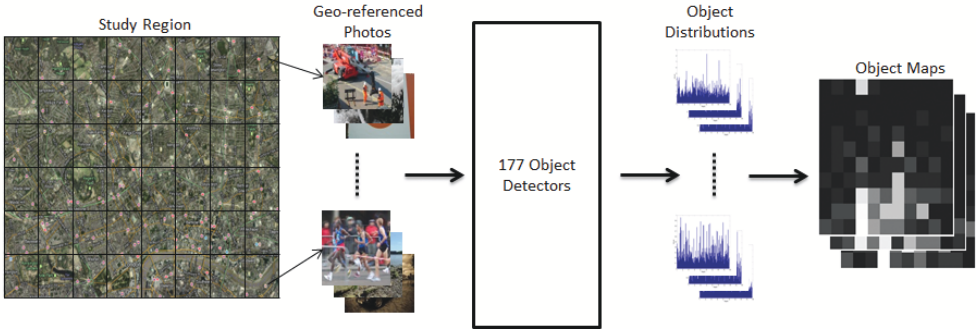


Figure 1: Framework for producing object maps.

Table 2: The 10 most heterogeneously distributed objects.

Objects	Homogeneity
Light	0.78
Sky	0.78
Fence	0.785
Desk	0.79
Gallery	0.79
Soil	0.795
Basketball hoop	0.8
Clock	0.8
Desktop computer	0.8
Boot	0.805

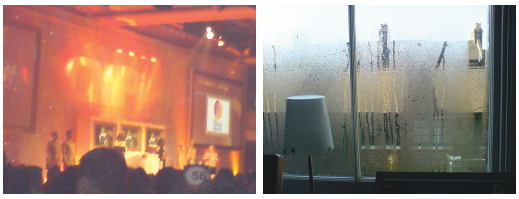
positives from the detections.

6. DISCUSSION

Our experimental results show promising opportunities of performing land-use classification by detecting objects and concepts from user contributed geo-referenced photos; challenges clearly remain however.

6.1 Noise in datasets

Although the object detectors we applied are considered to be state-of-the-art based on evaluation using standardized datasets in the computer vision community, they fail to perform as well in the real-life photo collections that contain many different types of photos and different styles of photography. This poses a challenge to using user-contributed photo collections for geographic knowledge discovery because many of these photos are not geographically informative. One aspect of our future work will focus on how to pre-process the photo collections so that non-useful photos will be removed from the collections before any image analysis takes place. One way of achieving this might be to employ image processing techniques to remove photos with poor image quality such as blurred and low-contrast photos. Furthermore, we can analyze the textual information accompanying the photos and discard photos without any geographically informative text.

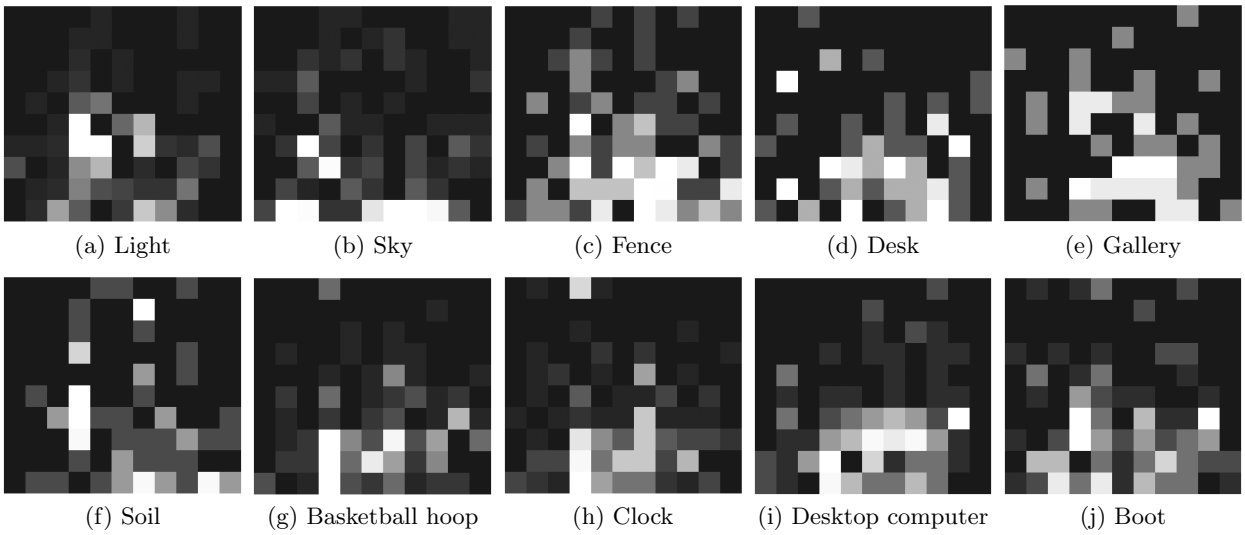


(a) (b)

Figure 2: Examples of false detections. (a) A basketball hoop is detected. (b) A boot is detected.

6.2 Latent information

Because the semantic information from the photo collections may not be extracted correctly due to the inaccuracy of the object detectors, we cannot determine the land-use class of any region directly based on the detected object appearances. However, the distinctiveness of the spatial distributions among objects suggests that the detectors are able to observe differences across the study region. Although the detected "objects" may not have any semantic meanings, they can serve as a mid-level, or latent, information that sits between low-level and high-level image analysis. In our future work, we will investigate the use of the resulting object distributions within each geographic tile as input features to perform land-use classification in a machine learning



(a) Light (b) Sky (c) Fence (d) Desk (e) Gallery

(f) Soil (g) Basketball hoop (h) Clock (i) Desktop computer (j) Boot

Figure 3: Spatial distributions of the 10 most heterogeneously distributed objects. Each block corresponds to a 1x1km region in the study area. The intensities of the blocks indicate the distribution of the detected objects.

framework.

7. CONCLUSIONS

In this work, we applied off-the-shelf object detectors to a collection of geo-referenced photos obtained from Flickr for the purpose of extracting semantic information from the collection. Although the detectors themselves have high detection errors, the maps they produce indicate a large range of spatial variation among objects and therefore may be used as a discriminative tool for land-use classification. In order to enhance the performance of the object detectors, further research on removing non-geographically informative photos from the collection is needed.

8. ACKNOWLEDGEMENT

This work was funded in part by an National Science Foundation CAREER grant (IIS-1150115) and a US Department of Energy Early Career Scientist and Engineer/PECASE award.

9. REFERENCES

[1] Flickr photo sharing. http://www.flickr.com.

[2] Picasa web albums. http://picasa.google.com/.

[3] D. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg. Mapping the world's photos. In *Proceedings of the International World Wide Web Conference*, pages 761–770, 2009.

[4] P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[5] D. Hoiem, A. A. Efros, and M. Hebert. Automatic photo pop-up. In *ACM SIGGRAPH 2005 Papers*, SIGGRAPH '05, pages 577–584, New York, NY, USA, 2005. ACM.

[6] D. Leung and S. Newsam. Proximate sensing using georeferenced community contributed photo collections. In *ACM International Conference on Advances in Geographic Information Systems: Workshop on Location Based Social Networks*, 2009.

[7] D. Leung and S. Newsam. Proximate sensing: Inferring what-is-where from georeferenced photo collections. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2955–2962, 2010.

[8] L. Li, H. Su, E. Xing, and L. Fei-Fei. Object bank: A high-level image representation for scene classification and semantic feature sparsification. In *Neural Information Processing Systems (NIPS)*, pages 1378–1386, Canada, 2010.

[9] J. Ponce, M. Hebert, C. Schmid, and A. Zisserman, editors. *Toward Category-Level Object Recognition*, volume 4170 of *Lecture Notes in Computer Science*. Springer, 2006.

[10] K. Yanai, K. Yaegashi, and B. Qiu. Detecting cultural differences using consumer-generated geotagged photos. In *Proceedings of the International Workshop on Location and the Web*, 2009.