Urban land use information retrieval based on scene classification of Google Street View images

Xiaojiang Li[1], Chuanrong Zhang[1]

[1]Department of Geography, University of Connecticut, Storrs
Email: {xiaojiang.li;chuanrong.zhang}@uconn.edu

Abstract

Land use maps are very important references for the urban planning and management. However, it is difficult and time-consuming to get high-resolution urban land use maps. In this study, we propose a new method to derive land use information at building block level based on machine learning and geo-tagged street-level imagery – Google Street View images. Several commonly used generic image features (GIST, HoG, and SIFT-Fisher) are used to represent street-level images of different cityscapes in a case study area of New York City. Machine learning is further used to categorize different images based on the calculated image features of different street-level images. Accuracy assessment results show that the method developed in this study is a promising method for land use mapping at building block level in future.

1. Introduction

Land use maps are very important references for urban planning and other urban practices in cities (Pei *et al*. 2014). Traditionally, overhead view remotely sensed data is widely used for land use/cover mapping based on different physical characteristics (spectral reflectance and texture) of different urban features (Pei *et al*. 2014). However, urban land use types are heterogeneous, and different land use types may have the same or similar spectral reflectance and spatial patterns. This makes it difficult to classify different land use types accurately based on remote sensing information alone. In addition, the remotely sensed imagery captures the roofs of buildings, which can hardly reflect the different social functions or land use types of buildings.

Different from the overhead view of remotely sensed imagery, Google Street View (GSV) images capture the profile view of streetscapes. GSV images have already been used to studying human perception of physical environment on ground (Li *et al*. 2015; Quercia *et al*. 2014). The street-level images represent the ground truth at a very high resolution and have been widely used as references for validating land cover/use mapping results manually in previous studies. The profile view street-level images could also be used to judge the land use types of different building blocks. However, based on our best knowledge, there still have no previous study using street-level images for urban land use mapping.

In the past decade, the advancement in the computer vision makes it possible to categorize and semantically classify images. In this study, we propose to bring scene classification algorithms in computer vision community to derive land use information of building blocks in cities based on geo-tagged GSV images. Multiple commonly used image features are calculated for representation of street-level images, which capture façades of different types of building blocks. Support vector machine classifier is then trained based on the calculated image features and ground truth land use labels and applied to predict land use types of different building blocks.

2. Data and Methods

2.1 Datasets

A small case study area in Brooklyn, New York City is chosen in this study. The study area includes various land use types, which is very suitable for testing the method using street-level images for land use mapping. Figure 1 shows the location and land use map of the study area.

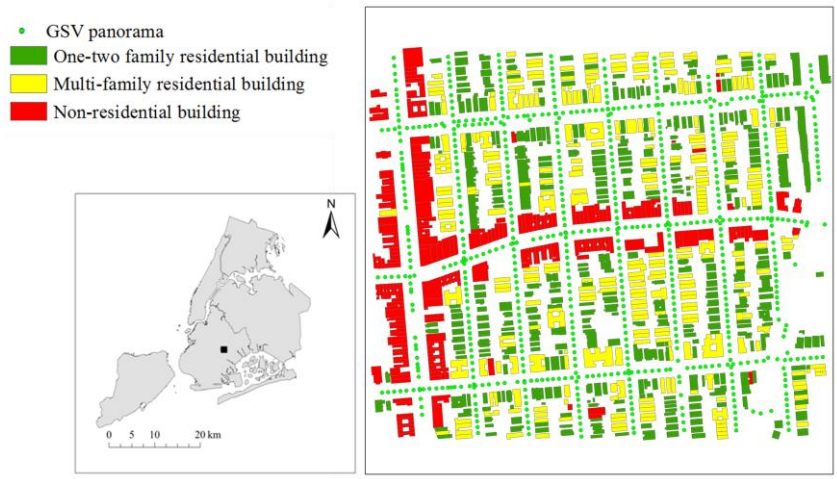

Figure 1. The location and land use map of the study area.

Figure 2. GSV images collections, (a) distributions of created sample sites and GSV panoramas, (b) static GSV images of one site with different heading angles, (c) geometrical model for choosing *heading* and *fov* to represent façades of building blocks.

2.2 GSV images collection and labelling

Google Street View panoramas are distributed discretely along the streets. In general, about every 12 meters has one GSV panorama along the street. Therefore, in this study we first create sample sites along streets every 5 meters using ArcGIS 10.2 in order to collect all available GSV panoramas along streets. We then retrieval the GSV panorama ID, coordinates information using Google Maps JavaScript API by inputting the coordinates of those sample sites. In this way, we collect the metadata (panorama ID and coordinate of panorama) of all available GSV panoramas along streets in the study area. Figure 2(a) shows the discrepancy of the distributions of created sample sites and location of GSV panoramas along streets in a small area of study area.

Based on the panorama ID information, we can download GSV images for different heading angles using Google Street View static Image API. Figure 2(b) shows four static GSV images of one site with panorama ID "*on66Bt1B37qRlYVxiC7J9g*" at different heading angles. By specifying appropriate *fov* and *heading* parameters, the GSV images can capture the façades of building blocks along streets, which makes it possible to differentiate different types of building blocks based on their different appearances. For each building block along a street, the closest GSV panorama is chosen. Based on the geometrical model between the location of GSV site (G_x, G_y) and the footprint of building block (see Figure 2(c)), we calculated the field of view (*fov*) angle by equation (1):

$$fov = \arctan\left(\frac{V_1 \cdot V_2}{\| V_1 \| \times \| V_2 \|}\right) \quad (1)$$

The vectors V_1 and V_2 are,

$$V_1 = (x_1 - G_x, y_1 - G_y),$$
$$V_2 = (x_2 - G_x, y_2 - G_y)$$

Where (x_1, y_1) and (x_2, y_2) are the coordinates of two endpoints of a building façade. In order to decrease distortion in the static GSV images, the *fov* cannot be too large, therefore, for those building blocks with *fov* larger than 90, the *fov* is set to 90. In addition, the minimum *fov* is set to 30 empirically, although some building blocks may have their *fov* less than 30. This is because if the *fov* is too small, the GSV image may not capture the spatial pattern of building block. The heading angle *heading* is set to the angle between heading direction and the true north direction, and ranges from 0 to 360. Figure 3 shows several collected GSV images with their corresponding land use types in the study area.

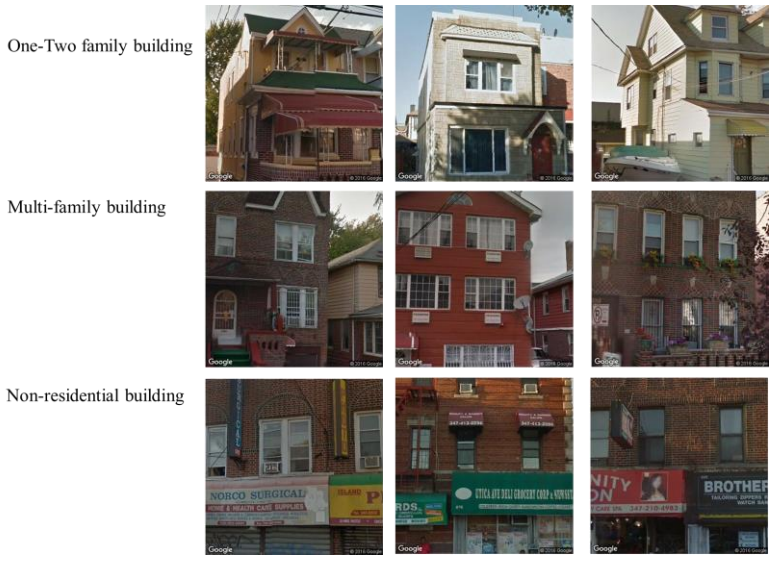

Figure 3. Building blocks with different land use types on street-level images.

2.3 Image features extraction and machine learning

The image features developed in computer vision community make it possible to represent and categorize street-level images of different cityscapes. Image features, which are calculated based on the texture and geometrical information of images, are insensitive to the variance of spectral information or the illumination conditions. In this study, several commonly used generic image features are used to indicate the characteristics of different street-level images. The image features used in this study include GIST, HoG, and SIFT-Fisher. Table 1 summarizes the descriptions of these image features. These features have already been tested on scene classification (Xiao *et al*. 2010) and semantic information retrieval from street-level images (Ordonez and Berg 2014; Naik *et al*. 2014). Therefore, in this study these features are chosen for representation of different GSV images and scene classification in terms of land use types.

Table 1. Image features used in this study for GSV image representation.

Image features	Descriptions
GIST	GIST is a 512 dimensional vector. The GIST feature is based on low dimensional representation of scene and represents the dominant spatial structure of a scene (Oliva and Torralba 2001).
HoG	Histogram of oriented edges (HoG) decomposes an image into small squared cells, computes a histogram of oriented gradients in each cell, normalizes the result using a block-wise pattern, and return a descriptor for each cell (Dalal and Triggs 2005).
SIFT-Fisher vectors	SIFT-Fisher vectors compute the SIFT features densely across five image resolutions, then perform spatial pooling by computing the Fisher vectors representations on a 2x2 grid over the image and for the whole image (Ordonez & Berg 2014; Perronnin *et al*. 2010).

To classify those static GSV images, which capture façades of different land use types of buildings, we choose Support Vector Machine (SVM) classifier. The original 1048 images and land use labels are split into training set and testing set. Image feature vectors $\boldsymbol{x}$ and their corresponding land use labels y in training set are used to train a SVM classifier. The training process is to obtain the following optimization,

$$\min \frac{1}{2} W^T W + C \sum_{i=1}^{l} \xi_i \qquad (2)$$

subject to

$$y_i (W \cdot x_i + b) \geq 1 - \xi_i, i = 1,2,..., N$$
$$\xi_i \geq 0, i = 1,2,..., N$$

where W is the support vector, ξ_i are slack variables introduced to account for the nonseparability of data, N is the number of training samples, constant C represents a penalty parameter that allows to control the penalty assigned to errors.

The trained SVM classifier is then applied to the testing images and compared with ground truth land use of these images in testing set to cross-validate the classification results.

3. Results

We collect 1048 static GSV images with different land use types in the study area. We randomly split these images into training set and test set 10 times to cross-validate the proposed method. Table 2 summarizes the cross-validation results using different image features. The SIFT-Fisher feature outperforms other two image features in the classification of residential building and non-residential building. The overall accuracy of the residential building *vs* non-residential building classification result is 91.82% using SIFT-Fisher image feature. The GIST and HoG features get lower classification results, with accuracy of 83.88% and 60.34% respectively.

The classification accuracy of one-two family residential building *vs* multi-family residential building has lower accuracy compared with the classification result of the residential building *vs* non-residential building. This is not difficult to understand, since the appearance difference between the one-two family residential buildings and multi-family buildings is not as obvious as the difference between the residential buildings and non-residential buildings. The selected three image features have similar performances in the classification of one-two family residential building *vs* multi-family residential building. SIFT-Fisher outperforms other two image features, with overall accuracy of 74%.

Table 2. Overall classification accuracy of different image features

Image features	Overall accuracy
	Residential building *vs* non-residential building
GIST	83.88%
HoG	60.34%
SIFT-Fisher	91.82%
	One-two family building *vs* multi-family building
GIST	66.72%
HoG	52.48%
SIFT-Fisher	74.30%

4. Conclusion and future works

This study brings scene classification algorithms in computer vision community to geospatial information retrieval based on publicly accessible data on the web. Different with previous studies using overhead view dataset for urban land use mapping, we first use street-level images, which capture the profile view of cityscapes, for land use classification at building block level. Accuracy assessment results show that using the combination of scene classification algorithms and street-level image is a very promising method for urban land use mapping. While this study demonstrates the feasibility of using GSV images for building block level land use information retrieval, there are still some limitations that need to be solved in the future studies. The basic idea of this study is to differentiate different types of land use types based on the different physical appearances of different types of buildings. However, the definition of different land use types is not based on the physical appearances of buildings, but the social functions of buildings. Therefore, in future studies, more attention need to be paid on how to make the semantic classification system to be applicable in real urban planning practices and theoretically recognizable at same time. Future work would focus on choosing better image features and combinations of image features to classify more land use types and get more accurate land use classification results. Human reasoning and new kinds of data should also been considered to get better classification results in future studies.

References

Dalal N and Triggs B, 2005, Histogram of oriented gradient object detection. In Proc. IEEE Conf. Computer Vision and Pattern Recognition, San Diego, USA 886-893.

Li X, Zhang C, and Li W, 2015. Does the Visibility of Greenery Increase Perceived Safety in Urban Areas? Evidence from the Place Pulse 1.0 Dataset. *ISPRS International Journal of Geo-Information*, 4(3), 1166-1183.

Ordonez V, and Berg T, 2014, Learning high-level judgments of urban perception, In Computer Vision–ECCV 2014, Springer International Publishing, 494-510.

Oliva A and Torralba A, 2001, Modelling the shape of the scene: A holistic representation of the spatial envelope, *International journal of computer vision*, 42(3), 145-175.

Pei T, Stanislav Sobolevsky, Ratti C, Shaw S, Li T, and Zhou C, 2014, A new insight into land use classification based on aggregated mobile phone data, *International Journal of Geographical Information Science* 28(9): 1988-2007.

Perronnin F, Sánchez J, & Mensink T, 2010, Improving the fisher kernel for large-scale image classification, In Computer Vision–ECCV 2010, Springer Berlin Heidelberg, 143-156.

Quercia D, O'Hare N, and Cramer H, 2014, Aesthetic capital: what makes London look beautiful, quiet, and happy? In Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing, 945-955.

Xiao J, Hays J, Ehinger K, Oliva A, and Torralba A, 2010, Sun database: Large-scale scene recognition from abbey to zoo, In Computer vision and pattern recognition (CVPR), 2010 IEEE conference on, San Francisco, USA, 3485-3492.