SEMANTIC CLASSIFICATION OF URBAN BUILDINGS COMBINING VHR IMAGES AND GIS DATA

S. Du[a]*, F. Zhang[a], X. Zhang[a]

[a] *Institute of Remote Sensing and GIS, Peking University, Beijing 100871, China*

KEY WORDS: Very high resolution (VHR) images, urban buildings, semantic classification, random forest, object-based image analysis (OBIA)

ABSTRACT:

While most existing studies have focused on extracting geometric information on buildings, only a few have concentrated on semantic information. The lack of semantic information cannot satisfy many demands on resolving environmental and social issues. This study presents an approach to semantically classify buildings into much finer categories than those of existing studies by learning random forest (RF) classifier from a large number of imbalanced samples with high-dimensional features. First, a two-level segmentation mechanism combining GIS and VHR image produces single image objects at a large scale and intra-object components at a small scale. Second, a semi-supervised method chooses a large number of unbiased samples by considering the spatial proximity and intra-cluster similarity of buildings. Third, two important improvements in RF classifier are made: a voting-distribution ranked rule for reducing the influences of imbalanced samples on classification accuracy and a feature importance measurement for evaluating each feature's contribution to the recognition of each category. Fourth, the semantic classification of urban buildings is practically conducted on two different kinds of cities, and the results demonstrate that the proposed approach is effective and accurate. The seven categories used in the study are finer than those in existing work and more helpful to studying many environmental and social problems.

1. INTRODUCTION

As main sites of urban activities and important components of cities, urban buildings are vital foundations of urban studies. Semantic classification of buildings intends to label buildings using a set of semantic categories cognized and conceptualized by people, such as low-story shantytowns, middle-story apartments, high-story apartments, administrative buildings, commercial buildings, etc. These categories strongly correlate with urban environment analyses (e.g. ecological and environmental evaluation), urban resource allocation (e.g. resource management, transportation planning, and disaster reduction) and urban social analyses (e.g. population estimation, and market research) (Wu et al., 2005). Existing work has focused on how to extract building contours or accurately distinguish buildings from non-buildings. However, geometric information alone cannot fulfill the demands on urban ecology, resources and social researches (Paul et al., 2001). Therefore, semantic classification of urban buildings is required.

Most existing studies have focused on extracting geometric information on buildings while only a few have concentrated on semantic analysis. In addition, some important issues still remain to be resolved. First, existing work on semantic analyses has distinguished too few categories to satisfy the many demands in environmental or social sciences (Graesser et al., 2012; Kohli et al., 2012; Belgiu et al., 2014; Lu et al., 2014). Second, there have been no appropriate segmentation scales and algorithms to produce single image objects for diverse buildings. This lack of computational methods leads to low classification accuracies as image features strongly depend on segmentation scales. Third, a small number of manually chosen samples and features may be practical for classifying a few categories of buildings. To distinguish more building categories greatly varying in sizes, shapes, structures, and spectrums, however, a large number of samples, high-dimension and heterogeneous features are required. In this situation, the samples are often imbalanced, and the features are often auto-correlated and have distinct importance for distinguishing different categories (Du et al., 2015). Unfortunately, there is still a lack of related work to reduce the influences of imbalanced samples on classification and to evaluate feature importance to classifying each category.

Aimed to resolve the issues raised above, this study presents a two-level segmentation mechanism (i.e. a large-scale layer constrained by GIS data for producing single image objects and a small-scale layer providing intra-object component features) and a semi-supervised method to choose a large number of unbiased samples by considering the spatial proximity and intra-cluster similarity of buildings. Random forest (RF) classifier is used to semantically classify buildings, for it is capable of handling a large number of samples and high-dimension and heterogeneous features. Moreover, to improve classification accuracy and evaluate feature importance, two improvements in RF classifier are presented: a voting-distribution-ranked rule for reducing the influences of imbalanced samples and a feature importance measurement for each category based on Gini descent and path tracing strategy.

2. METHODOLOGY

2.1 Category system of urban buildings

Urban buildings made of various materials with assorted styles and appearances in the real physical world are the basis of cognizing semantic category by people and of sensing buildings by remote sensors. In the geoinformatic world, buildings are abstracted into contours in GIS data and into image pixels or

* Corresponding author

image objects in VHR images. Accordingly, they are described from the aspects of spectrum, shapes, and textures. In the cognition world, people cognize, understand, and communicate their ideas about buildings through appropriate semantic categories. Therefore, building a semantic category system helps to transform the feature representations in the geoinformatic world to the concepts in the cognition world.

The goal of semantic classification is to build relationships between the concepts of buildings in the cognition world and the features of buildings in the geoinformatic world. Therefore, the semantic category system can be built by discriminating the appearances and functions of urban buildings, including low-story (LS) shantytowns, medium-story (MS) apartments, high-story (HR) apartments, administrative (AD) buildings, commercial (CM) buildings, industrial (ID) buildings, and auxiliary (AU) buildings.

2.2 The presented procedure

To solve the issues raised, such as the complete segmentation of buildings, the choice of training samples, and the classification approach of semantic categories, this study presents an approach to semantic classification of urban buildings.

Four steps are required.

(1) **VHR image segmentation constrained by GIS data**. Since existing segmentation methods cannot produce single image object for each building, this study adopts a segmentation constrained by GIS data to obtain a single object for each building. The pixels inside each GIS contour are first merged into a single object, and then image features are computed.

(2) **Features extraction of buildings.** Image features are the bridge connecting building objects to semantic categories. In this study, four types of features are used, including spectrum, texture, geometry, and spatial distribution. The first three choose samples and classify buildings while the last evaluates clustering results and chooses samples.

(3) **Samples selection with semi-supervised method**. A large number of samples is required to semantic classification. The ISODATA algorithm is first used to cluster buildings according to spectrum, geometry, and texture features. Then, intra-cluster similarity is used to choose corresponding samples of semantic categories in a semi-supervised way.

(4) **Semantic classification of buildings using improved RF classifier.** The three steps above can collect a larger number of samples as well as high-dimension and heterogeneous features for each sample. To train a classifier from these complex features and huge samples, a semantic classification approach using RF classifier is presented. Furthermore, RF is improved to evaluate the contribution of each feature to each category, which reduces the influences of imbalanced samples.

2.3 Improved RF classifier for semantic classification

To reasonably exploit samples and their features, the RF classifier is employed to randomly select samples and features to train decision trees and to integrate all trained decision trees to vote for the most popular category (Breiman, 2000). In this study, RF is improved to evaluate the contribution of each feature to classifying each category and to reduce the influences of imbalanced samples on classification results.

Based on bagging integration learning algorithm, RF classifier trains each decision tree independently (Pal, 2005) and the votes of all decision trees determine the final results. The training steps are as follows: (1) to choose a subset of samples using Bootstrap sampling methods, (2) to choose randomly $\sqrt{M}$ features from M ones for each node, (3) to construct a CART decision tree with the chosen samples by using GINI coefficient as information gain, and (4) to build N CART decision trees until a RF is built.

The classification process of decision trees is exactly same as that of the training process. For each building, each tree independently predicts a category; accordingly, the resulting category is the most popular category.

Traditional RF uses the simple majority voting rule to make decisions and tends to misclassify the minority categories. Therefore, imbalanced samples heavily affect the classification accuracy. Existing work on learning imbalanced data includes pseudo balanced random forest (BRF) and weighted random forest (WRF) (Chen et al., 2004). BRF copies a small number of samples for the minority categories and randomly chooses the same number of samples for the majority categories. WRF weighs the samples of different categories to reduce the differences in samples. However, BRF needs to identify adaptively the number of copied samples, causing it to change the original distribution of samples. Meanwhile, WRF necessitates the specification of weight for each category. However, how to specify weights is still unresolved. Therefore, a new voting rule - the voting-distribution ranked rule - is proposed to replace the simple majority voting rule.

Supposing there are N decision trees and K categories, and each tree has one vote for a sample (o), then the votes of the N trees can be represented as a distribution $Vote(o) = (n_1, n_2, \dots, n_K)$ with $\sum_{i=1}^{K} n_i = N$. The simple majority vote rule (Kontschieder et al., 2014) assigns a sample to the most popular category.

The vote distributions of out-of-bag (OOB) samples are significant in discovering confused categories. For example, let n_1 and n_2 be the first and second maximum votes; if they are close, unclassified samples tend to be misclassified. Due to the influence of imbalanced samples, it is easier for majority categories than minority categories to obtain more votes. Once the votes of the majority categories are larger than those that they deserved, misclassification will occur. Fortunately, imbalanced samples do not hide the vote distribution over categories, even though they affect the number of votes for each category. Let $p_i = n_i/N \ \ (i = 1, 2, \dots, K)$ be the probability of the i *-th* category, then the vote distribution of unclassified sample o is defined as $prob(o) = (p_1, p_2, \dots, p_K)$. For each OOB sample, a vote distribution can be obtained; as a result, multiple vote distributions can be obtained for each category since there are many OOB samples for each category. The reliable distributions (ranked in top 5% in the probabilistic distributions) of each category are averaged into a representative one. Accordingly, for an unclassified sample, its vote distribution is compared with the representative distributions of the K categories. The sample is then assigned to the category with the largest similarity.

3. EXPREMENTAL RESUTLS

To verify the feasibility of the presented approach, a series of experiments are conducted, including sample selection, semantic classification of traditional RF, semantic classification of improved RF for handling imbalanced samples, and evaluation of feature importance.

3.1 Study area and used data

Since urban buildings vary notably in sizes, shapes, structures and spectrums, the experiments will focus on them to test the presented method. The study area is located at Haidian and

Xicheng districts in Beijing city (Fig. 1), which belongs to the city expansion area and is full of high-tech companies, culture and education industry, commercial and service industry, and research institutions. In addition, due to rapid urbanization in recent years, the area has a large number of informal settlements and developing zones. Therefore, it is very significant to analyze and address the semantic categories of buildings in this area.

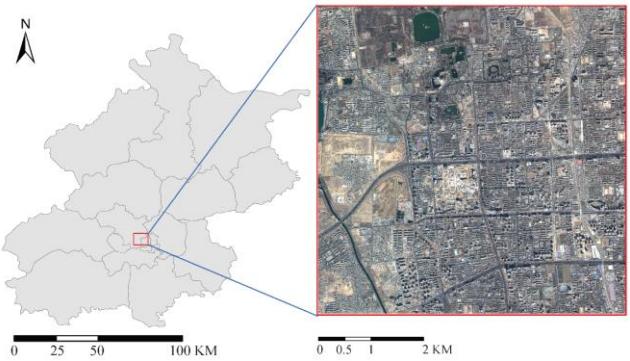

Fig. 1. Study area and Quickbird data.

For semantic classification of buildings, both GIS and Quickbird data are used:

(1) **Quickbird image Data**. Panchromatic band with resolution of 0.61m and four multispectral bands with resolution of 2.44m are fused to produce a four band data with resolution of 0.61m. The spectral, geometric, and texture features are extracted from Quickbird image.

(2) **GIS data of buildings**. To obtain a single-image object for each building, the contours in GIS data were used. The study area covers 38.7 km^2 and contains 8831 buildings. The largest contour, about 96462 m^2, represents a LS shantytown or ID building, while the smallest contour is only 70 m^2 and refers to an AU building.

Before semantic classification and sample selection, two-level segmentation was conducted and 307 image features were computed for each image object. In total, 8831 image objects and 15258 sub-objects were obtained. The largest object has 118 sub-objects while the smallest object has only one sub-object.

3.2 Classification results with the simple majority vote rule

A RF classifier with 200 decision trees is trained using the chosen 2747 samples, and the 6084 unclassified buildings are classified using the trained RF. For each OOB sample, each decision tree will have one vote for classification. The final category is the one with the most votes, and the confusion matrix of accuracy assessment is created based on the predicted and existing categories of OOB samples. The overall accuracy is 71.50%, and overall kappa coefficient 0.59. The overall accuracy is greatly reduced by the misclassification of 303 high-story apartments and 35 commercial buildings.

3.3 Classification results with the voting-distribution ranked rule

The imbalanced samples result from the real distribution of buildings in each category, and they will lead to a certain bias when identifying the categories using the simple majority voting rule. The minority categories will probably be misclassified. Due to the small number of samples, MS apartments - as are CM and AD buildings - are often confused with HR apartments. This confusion makes it difficult to correctly classify these buildings by the simple majority voting rule. The voting-distribution ranked rule (Du et al., 2015) is used to reclassify unclassified buildings (Fig. 2). The overall accuracy increases to 79.54%, and the kappa coefficient remarkably to 0.72, demonstrating that our new vote rule can improve the classification accuracy: the misclassified MS and HR apartments are greatly reduced.

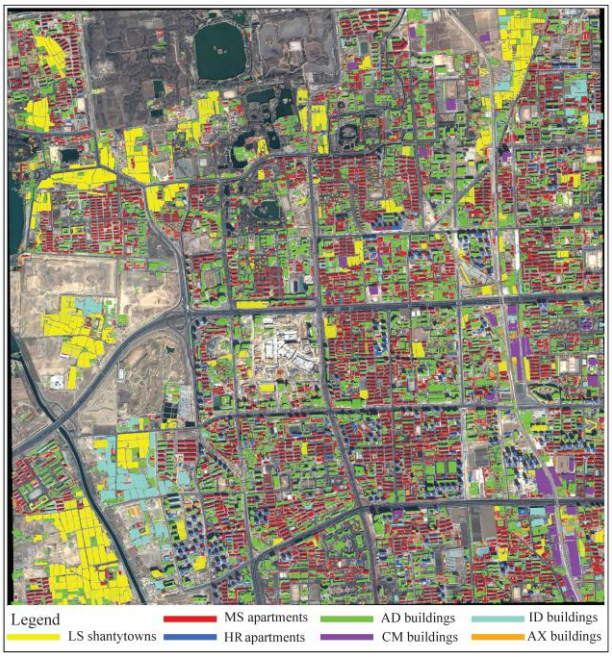

Fig. 2. The classification results with the improved voting rule.

3.4 Evaluation of feature importance

The image features used in this study are highly correlated, high-dimensional and heterogeneous; therefore, it is necessary to evaluate the contributions of these features to classification. Based on the trained RF, the feature importance analysis method (Gini descent) was used to evaluate the importance of the 307 features.

In Table 1, texture and geometric features have great contributions to most categories, while spectral ones have poor performances due to their confusions. LS shantytowns are much larger than other buildings, and thus geometric features (e.g. area and perimeter) are significant. In addition, AU buildings are usually relatively small and easily distinguished by geometric features. Geometric features also play important roles in classifying MS apartments and ID and CM buildings because of the buildings' unique shapes and areas; however, HR apartments often have complex roofs and structures, showing unique characteristics of texture, thus they can be identified by texture features. Composed of diverse sub-objects, AD buildings vary greatly in materials and styles; thus, they can be classified by combining geometric and texture features. In other words, geometric and texture features are more important to classification than spectral ones.

Table 1. The feature importance scores to seven categories.

Semantic category	Feature contribution rank (%)		
	spectrum	geometry	texture
LS shantytowns	15.36	31.80	52.84
MS apartments	12.30	45.14	42.56
HR apartments	25.94	15.37	58.69
AD buildings	10.71	40.27	49.02
CM buildings	3.39	55.00	41.62
ID buildings	15.27	38.28	46.45
AU buildings	15.80	63.07	21.13

4. CONCLUSION

This study presents an improved classification approach for semantic classification of urban buildings. Four scientific tasks are resolved. Initially, GIS data were used to constrain the image segmentation for producing a single-image object for each building. Then, at the second level, each image object is further split into sub-objects to measure the internal heterogeneity of buildings. Next, ISODATA algorithm was used to group image objects into clusters by using extracted features, and a large number of unbiased samples were chosen by considering spatial proximity and intra-cluster similarity. The chosen samples reflect the real distributions of buildings in the physical world. Subsequently, the voting-distribution ranked rule can improve RF classifier by reducing classification error caused by imbalanced samples. The classification results of the improved and original RF were compared, and the accuracy increased from 71.50% to 79.54%.

REFERENCES

Belgiu, M., Tomljenovic, I., Lampoltshammer, et al., 2014. Ontology-based classification of building types detected from airborne laser scanning data. *Remote Sensing*, 6, pp. 1347-1366.

Breiman, L., 2000. Randomizing outputs to increase prediction accuracy. *Machine Learning*, 40(3), pp. 229-242.

Chen, C., Liaw, A., Breiman, L., 2004. Using random forest to learn imbalanced data. Technical Report of Department of Statistics, UC, Berkeley.

Du, S., Zhang, F., Zhang, X., 2015. Semantic classification of urban buildings combining VHR image and vector data: a random forest approach. *ISPRS Journal of Photogrammetry and Remote Sensing*, 105, pp. 107-109.

Graesser, J., Cheriyadat A., Vatsavai R. R., Chandola, V., Long, J., and Bright E., 2012. Image based characterization of formal and informal neighborhoods in an urban landscape. IEEE Journal of selected topics in applied earth observations and remote sensing, 5(4), pp. 1164-1176.

Kohli, D., Sliuzas, R., Kerle, N., Stein, A., 2012. An ontology of slums for image-based classification. Computers, Environment and Urban Systems, 36, pp. 154-163.

Lu, Z., Im, J., Rhee, J., Hodgson, M., 2014. Building type classification using spatial and landscape attributes derived from LIDAR remote sensing data. Landscape and Urban Planning, 130, pp. 134-148.

Pal, M., 2005. Random forest classifier for remote sensing classification. International Journal of Remote Sensing, 26(1), pp. 217-222.

Paul, S., et al., 2001. Census from heaven: an estimate of the global human population using night-time satellite imagery. International Journal of Remote Sensing, 22(16), pp. 3061-3076.

Wu, S., Qiu, X., and Wang, L., 2005. Population estimation methods in GIS and remote sensing: a review. GIScience & Remote Sensing, 42(1), pp. 80-96.