Detection of Urban Zones in Satellite Images Using Visual Words

Lior Weizman and Jacob Goldberger

Abstract—Today, satellite and aerial images are the major source of information for landcover classification. An important usage of remotely sensed data is extracting urban regions to update GIS databases. However, in most cases human resources do not give a sufficient solution to the problem, since it can not entirely process such an enormous amount of remotely sensed data. In addition, most of the automatic methods for urban extraction that exist today are sensitive to atmospheric and radiometric parameters of the acquired image. In this paper we address the problem of urban areas extraction by using a visual representation concept known as “Bag of Words”. This method, originally developed for text retrieval approaches, has been successfully applied to scenery image classification tasks. In this paper we introduce the “Bag of Words” approach into analysis of aerial and satellite images. Due to the fact that we implement a normalization process in our method, it is robust to changes in atmospheric conditions during acquisition time. The improved performance of the proposed method is demonstrated on IKONOS images. To assess the robustness of our method, the learning and testing procedures are performed on two different and independent images.

Keywords - Classification, object detection, remote sensing, multispectral.

I. INTRODUCTION

In the last few years, urban zone detection from satellite sensor imagery has become important for several purposes. The main purpose is Geographic Information Systems (GIS) update, a continuous need which enables an efficient study and planning of urban growth. In addition, this information may help government agencies and other policy makers arriving at decisions about regions. In most cases, humans are not a sufficient resource to handle the enormous number of satellite images acquired for urban detection. Therefore, it is essential to have efficient tools for automatic detection and segmentation of urban areas. Because of the unique texture of urban scenes with respect to natural scenes, the major approaches for segmentation of urban zones are based on texture analysis. The texture operators can be generally divided into gray-level-based and structure-based texture operators. The usual gray-level-based texture operators include gray-level co-occurrence matrix (GLCM) [7], normalized gray-level histogram [6], gray-level difference, and gray-level run length [9]. Among the structure-based operators we can find: Gabor wavelet [3], gradient-based feature [10] and Markov Random Field [4]. Recent works by Zhong and Wang [11], [12] combine low and high level of structure-based texture for urban detection. Although very different in approach, all the currently used methods for urban detection suffer from a major drawback - the absence of robustness. Most of satellite

imaging (even when acquired by the same sensor above the same area) would have different gray-levels scale due to the atmospheric conditions during the capturing time. Algorithms for atmospheric correction are mostly time consuming and rarely overcome the above mentioned problem.

This work presents a new approach for the task of urban detection and segmentation. The method is based on the “Bag of Words” (BoW) paradigm which is a recently introduced concept that has been successfully applied to scenery image classification tasks (see e.g. [2], [1]). The BoW model is based on the idea that it is possible to transform the image into a set of visual words and to represent the image (and objects within the image) using the statistics of appearance of each word as feature vectors. The visual words are image patches (small sub images) that are clustered to form a dictionary consisting of a small set of representative patches. We utilize this approach for the purpose of urban areas extraction, while implementing modifications which are relevant for the urban segmentation task. Therefore, our method is robust to changes in scene and to atmospheric effects.

The rest of this paper is organized as follows. The theoretical development of the method is given in Section II. Experimental results on IKONOS images are presented in Section III. Section IV summarizes the paper with discussions and conclusions.

II. THE URBAN DETECTION ALGORITHM

In this study we show that highly successful text retrieval approaches (known as “Bag-of-Words”) can be used for detecting urban areas in satellite images. In this model, a text (such as a sentence or a document) is represented as an unordered collection of words, disregarding grammar and even word order. In this view of a document we only retain information on the number of occurrences of each word. In the bag-of-words model a document is statistically modeled as an instance of a multinomial word distribution and is represented as a frequency of occurrence word histogram (see [5] for an excellent introduction to the topic).

To represent an image using BoW model, an image has to be treated as a document. This requires to define a visual analogy for a word and a visual analogy for a code-book or a dictionary that contains a list of all possible words. In our terminology, a small patch of an image is defined as visual word. In the first step of our urban detection systems, a dictionary of visual words is built. In the next step we build visual words histograms for urban and non-urban areas. As a result, a set of “urban words” is defined. These words

occur much more frequently in urban areas and detection of such word is a strong indication of an urban region. Given a new unlabeled test image we look for visual-words that correspond to urban detection words as a first step for detecting urban areas. A postprocessing step applies spatial consistency constraints on the detected urban patches to obtain a global decision on urban regions. Following is a detailed description of the urban detection algorithm.

A. Building a dictionary

The task of forming a visual dictionary is the process of creating a vocabulary of words that will be further used to represent primitives in the image. In order to build a comprehensive dictionary, one or more images with urban and non-urban areas are required. We represent each image as a collection of spatially adjacent pixels (patches) which are treated collectively as a single primitive. We view patches of size $n \times n$ as one-dimensional vectors of size n^2 . To increase the robustness of the algorithm and to avoid the need for atmospheric/radiometric calibration, each vector is first normalized by subtracting the patch mean. This makes the features invariant to gray-level scale differences between images.

To reduce both the algorithm computational complexity and the level of noise, a feature extraction method is applied. Generally, urban zones, in contrast to non-urban zones, are characterized by high spatial frequencies. Therefore, we apply principal component analysis procedure (PCA) in order to reduce the dimensionality of the data. We expect that the first components of the PCA (which are the components with the highest variance in the image) will contain the information about the spatial frequencies of the patch, and therefore, will best differentiate urban zones from non-urban zones. Let $v_1, \dots, v_k$ be the k eigen vectors of the data covariance matrix corresponding to the k largest eigen values $\lambda_1, \dots, \lambda_k$. Applying the PCA transformation we project the original data into a k -dimensional subspace using the transformation defined by the matrix $(v_1/\sqrt{\lambda_1}, \dots, v_k/\sqrt{\lambda_k})^T$. The data variance in the projected space is the identity matrix.

The main step in the dictionary building procedure is clustering the patches to form a small-size dictionary of visual words. A common clustering algorithm, such as iterative self organizing data analysis (ISODATA) [8] or K-means can be used for this purpose. As a result, the data vectors in the projected space are clustered into M groups. Finally, the mean vector of every group is computed to create a dictionary with M visual words. Note that this dictionary building step is done in an unsupervised mode without any reference to the urban/non-urban label of each patch.

B. Urban words learning phase

In this stage, the relevant words from the dictionary that best differentiate urban areas from non-urban areas are found. First, urban and non urban areas are defined on the training image. Each area is then divided into patches and a mean normalization is performed on every patch following by the linear PCA transformation (that was computed in the previous

step) and assignment of the patch to the nearest dictionary word (using Euclidian distance). As a result we obtain two word frequency histograms, one of urban zone and one for non-urban zone. These histograms represent the amount of usage of every word from the dictionary in the urban and non urban areas. Normalizing the histograms we can view them as discrete distributions $P_{\text{urban}}(\cdot)$ and $P_{\text{non-urban}}(\cdot)$ of the visual words in urban and non-urban areas respectively.

Our goal is to find the words in the dictionary whose usage in urban areas is significantly higher than their usage in non-urban areas. Therefore, given an arbitrary patch, the probability of this patch being taken from an urban region can be computed using the Bayes rule:

$$P(\text{urban}|u) = \frac{\alpha P_{\text{urban}}(u)}{\alpha P_{\text{urban}}(u) + (1 - \alpha) P_{\text{non-urban}}(u)} \quad (1)$$

where α is the prior probability of a patch to be in an urban region. The words from the dictionary that best differentiate urban areas from non-urban area are the words that satisfy the following inequality:

$$P(\text{urban}|u) \geq \text{threshold} \quad (2)$$

while the *threshold* is a tunable parameter. As a result we obtain a group of “urban words” that characterize urban patches. Detection of such word is a strong indication of an urban region.

C. Urban detection in a new image

Given a new image we want to detect and segment the urban regions. Each one of the image patches on a regular grid is translated to one of the visual words from the dictionary. This is done by first normalizing the patch vector, applying the PCA transformation that was learned in the training step. Then, every transformed vector word is assigned to its nearest word from the dictionary (based on Euclidian distance). Utilizing equation (1), we can compute for each patch the posterior probability to be in a urban region. The result is a local urban/non-urban decision for each separate patch. To remove outliers and to obtain a global smooth decision on urban regions, a post-processing morphological operator is applied on the local urban-decision map. The majority voting analysis, in order to replace “holes” in the urban detection areas with their surroundings value, was found to be sufficient to achieve reliable global smooth results.

III. EXPERIMENTAL RESULTS

This section presents the results of the proposed method when applied to real satellite images. The experiments are based on two IKONOS images: the first one is used for building the dictionary and for the urban-words learning processes; the performance of the method is evaluated on the second image. The first image, presented in Fig. 1, was acquired on August 2005 and the second on April 2007. The spatial dimensions of the images are 4186×3054 and 3783×3010 , respectively.

Fig. 1. The training image.

Fig. 2. The dictionary of 60 words. The word's numbering order is left to right, one row after another.

A. Building a dictionary

In our implementation, the training image was divided into patches of size 10×10 each. As a result, 127490 patches were obtained. Normalization of every patch by subtracting the patch mean was then performed. The process was followed by a dimensionality reduction step to reduce the data to a dimension of 10. Applying the PCA, every patch in the image is approximated by a linear combination of these 10 patches. As explained in the previous section, the next step in the dictionary building process is to cluster the reduced data into M groups. We used the ISODATA clustering method to cluster the projected vectors into a dictionary of $M = 60$ words. Fig. 2 shows the words in the dictionary.

B. Urban words learning phase

In this phase, the relevant words that best differentiate urban areas from non urban areas are found. First, urban and non-urban areas were defined on the training image. Then, the frequencies of every word in the dictionary in the urban and non-urban areas were computed. We found that more than 95% of the non-urban area is modeled by a single word (word #2), while the histogram of the urban area is spread over the dictionary words. This phenomenon is explained by the fact that non-urban areas have low intra-variability, and therefore most of the non-urban patches in the image are grouped into

Fig. 3. The posterior probability of every word to be part of an urban scene.

a single patch during the clustering process. The gray level difference among non-urban patches is eliminated during the patch mean normalization step. On the other hand, urban areas, that are characterized by high variability, are modeled by the majority of the words in the dictionary.

The next step in the learning process is to find the words that have the highest posterior probability to be part of an urban scene, according to the probabilistic model that is defined in the previous section. We defined the prior probabilities as 0.5 each. The posterior probability of every word in the dictionary to be part of an urban scene, which was computed according to eq. (1), is presented in Fig. 3.

The final step in the learning process is defining the “Urban words” set, the words that their posterior probability to be part of an urban scene is above a certain threshold. We set the threshold to be $threshold = 0.95$. As a result, 43 words compose the “urban words” set. The indices of these words in the dictionary presented in Fig. 2 are: 1, 3 – 21, 24, 28 – 30, 32, 34 – 36, 38 – 41, 44 – 47, 49 – 53, 57 and 58. It can clearly be seen that most of the words that are included in the “Urban words” set exhibit morphological features that mostly characterize urban scene (e.g. edges, corners), while most of the words that are not included in the “Urban words” (e.g. 2, 25 and 37) do not include these features.

C. Urban detection in a new image

In order to detect urban areas in a new image, the test image was divided into patches of 10×10 each. As a result, a total number of 113778 patches were obtained. The same pre-processing was applied to assign a visual word for each patch. Finally, only the patches that were assigned to words that are included in the “Urban words” set were classified as urban patches. Morphological operator was then applied to the classification results image in order to remove outliers and to impose smoothness. We used a majority vote analysis with a kernel size that is 5 times larger than the patch spatial size, in order to fill the “holes” in the urban detection results.

In order to quantify the results of urban areas detection in a new image, urban areas were defined in the test image to

Fig. 4. Detection results (top) vs. ground truth (bottom) of urban areas in the test image.

create a ground truth image. A total number of 29 urban areas were defined on the ground truth image. Two quality measures were used. The probability of detection (PD) is defined as the number of urban areas that were detected in the test image divided by the total number of urban areas in the image (29). The false alarm rate (FAR) is defined as the number of pixels the were false detected as urban areas, divided by the total number of pixels in the image. The urban detection results vs. the ground truth image is presented in Fig. 4. In the exhibited results, 28 out of 29 urban zones were detected, (PD=95.6%) while the FAR is 2.19%.

IV. DISCUSSIONS AND CONCLUSIONS

In this study we have proposed a method to learn and recognize urban areas in satellite images. The common methods that exist today for urban area detection relies on extracted features from the images. Our proposed method makes the feature extraction procedure, which is mostly a time consuming procedure, unnecessary. Therefore, the BoW method has also a major advantage over the other methods for urban area detection, since it is not constrained to extract an arbitrary set of features, and therefore its robustness to changes in scene and atmospheric conditions is higher. The detection results

presented here convincingly demonstrate the power of using the BoW model for detection of urban zones. Our model requires prior learning of the distribution of the dictionary words usage in urban zones. Indeed, there are common detection and classification methods that also require prior learning, but our method is unique in its robustness and exhibits reliable results. This is despite the fact that the learning and test images are different in both the acquired scene and the acquisition time. However, a few trade-offs regarding the optimal selection of the parameters of the method have to be addressed. Currently, our framework still depends on optimal selection of the method's parameters in order to achieve the best results. Even so, we believe that the set of parameters that exhibited reliable results for a certain sensor, would exhibit similar results when applied to images acquired by the same sensor. In further work we will examine the performance of the method on a large database of images that include several types of sensors. Another topic for future research is to investigate the use of other dimensionality reduction techniques and their affect on the detection results. In Addition, we believe that this method can also be successfully used for other detection or classifications purposes of remotely sensed data. Therefore, another major contribution of this paper is simply introducing the BoW method to the remote sensing community.

REFERENCES

- [1] L. Fei-Fei and P. Perona, "A Bayesian hierarchical model for learning natural scene categories," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, pp. 524-531, 2005.
- [2] R. Fergus, P. Perona and A. Zisserman, "Object class recognition by unsupervised scale-invariant learning," *IEEE Conference on Computer Vision and Pattern Recognition (ICCV)*, vol. 2, pp. 264-271, 2003.
- [3] J. Li and R. M. Narayanan, "Integrated spectral and spatial information mining in remote sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 42, no. 3, pp. 673-685, Mar. 2004.
- [4] A. Lorette, X. Descombes, and J. Zerubia, "Texture analysis through a Markovian modelling and fuzzy classification: Application to urban area extraction from satellite images," *International Journal of Computer Vision*, vol. 36, no. 3, pp. 221-236, 2000.
- [5] C. D. Manning, P. Raghavan and H. Schutze, *Introduction to information retrieval*, Cambridge University Press, 2008.
- [6] A. K. Shackelford and C. H. Davis, "A hierarchical fuzzy classification approach for high-resolution multispectral data over urban areas," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 41, no. 9, pp. 1920-1932, Sep. 2003.
- [7] P. C. Smits and A. Annoni, "Updating land-cover maps by using texture information from very high-resolution space-borne imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 37, no. 3, pp. 1244-1254, May 1999.
- [8] J. T. Tou and R.C. Gonzalez, *Pattern Recognition Principles*, Addison-Wesley, 1977.
- [9] J. S. Weszka, C. R. Dyer, and A. Rosenfeld, "A comparative study of texture measures for terrain classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. SMC-6, no. 4, pp. 269-285, Apr. 1976.
- [10] S. Yu, M. Berthod, and G. Giraudon, "Toward robust analysis of satellite images using map information-Application to urban area detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 37, no. 4, pp. 1925-1939, Jul. 1999.
- [11] P. Zhong and R. Wang, "Using combination of statistical models and multilevel structural information for detecting urban areas from a single gray-level image," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 5, pp. 1469-1482, May 2007.
- [12] P. Zhong and R. Wang, "A multiple conditional random fields ensemble model for urban area detection in remote sensing optical images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 12, pp. 3978-3988, Dec. 2007.