INTRA-URBAN LAND COVER CLASSIFICATION FROM HIGH-RESOLUTION IMAGES USING THE C4.5 ALGORITHM

C. M. D. Pinho [a], F. C. Silva [a], L. M. C. Fonseca *, A. M. V. Monteiro[a]

[a]Image Processing Division , Brazil's National Institute for Space Research 1758 Astronautas, São José dos Campos, Brazil - (carolina, felipe,leila, miguel)@dpi.inpe.br

KEY WORDS: Remote Sensing, Land Cover, Classification, Image Analysis, Data Mining, Knowledge Base, High resolution.

ABSTRACT:

Nowadays, the availability of high-resolution images has increased the number of researches on urban land use and land cover classification. Most of them have used object oriented image analysis with successful results. Although object oriented analysis offers effective tools to represent the knowledge of the scene, the tasks of building semantic network and selecting attributes are time-consuming. These processes demand considerable prior knowledge of the scene and of the urban object characteristics. Therefore, we propose to use the C4.5 decision tree algorithm to help semantic network construction and attribute selection processes. This algorithm selects the best subset of attributes based on an entropy measure and organizes the classes in a decision tree structure. To evaluate the performance of C4.5 algorithm, we conduced a land cover classification in an urban area of São José dos Campos (São Paulo state, Brazil). Two experiments were performed, one based on specialist knowledge using E-Cognition 4.1™ system and the other based on the decision tree generated by C4.5 algorithm. Both provided similar results although the C4.5 experiment was faster than the other.

1. INTRODUCTION

The object-based image analysis has been a well-established method for the high-resolution image classification, mainly in urban areas (Centeno, Miqueles, 2004; Herold et al., 2003; Hoffman, 2001; Pinho, 2004; Rego, 2003; Thomas, 2003). In this approach the specialist models his/her knowledge through a semantic network. The multiresolution segmentation process extracts image objects. Afterwards, a set of spectral, spatial and geometrical attributes (more than a hundred) is calculated for each object. These processes integrated in the same framework provide a flexible and effective environment for the specialist to formalize the knowledge of the scene.

The e-Cognition software (Definiens, 2003) uses the object oriented image analysis for image classification and has been largely used by the remote sensing community. Nevertheless, the great number of object attributes and the different ways to model the semantic network make the task of classification lengthy and complex.

On the other hand, the use of data mining techniques can solve this problem because they quickly select the most representative attributes for each class and generate simple classification rules. Some remote sensing applications have successful used data mining techniques for different tasks: to retrieve specific spatial patterns (Silva et. al, 2007), to make land cover classification (Aksoy et al., 2004), to characterize and extract features (Datcu et al., 2003).

In this work, we propose to use a data mining technique to help attribute selection and semantic network construction tasks. The choice of decision tree algorithms as data mining tool (Aksoy et al., 2004) was due to the following factors: (1) little processing time is demanded; (2) the model is easily understood; (3) representative attributes are easily identified; (4) classification rules are simple; (5) object attributes can be represented as numerical and categorical. Besides, decision tree algorithms do not need any assumptions about either the statistical distributions or the independence of the classes.

We have chosen the C4.5 decision tree algorithm because it is freely available (on software WEKA) and has produced good results in remote sensing imagery classification (Silva, 2005). To test the methodology proposed in this paper, two experiments were performed. One based on specialist knowledge using E-Cognition 4.1™ system and the other based on the decision tree generated by C4.5 algorithm. The tests were carried out in a small area of São José dos Campos city (São Paulo state, Brazil). The next sections describe the study area and material. Subsequently, the methodology and results are presented.

2. THE STUDY AREA AND MATERIAL

The municipality of São José dos Campos is in the southeastern State of São Paulo (figure 1), within coordinates 46º06'W, 23º18'S, 45º40'W and 22º49'S. Its total surface amounts to 1,099.60 km2 and it has a population of about 539,313 inhabitants, out of which 532,717 (98.78 %) inhabitants live in urban areas. The municipality is along the road that connects São Paulo to Rio, and is famous for being the country's aerospace pole.

In the experiments, only a part of the south of the city was processed. This selected study area, although reduced in size as opposed to the whole urban area of São José dos Campos, contains a great diversity of intra-urban land cover classes.

* Corresponding author.

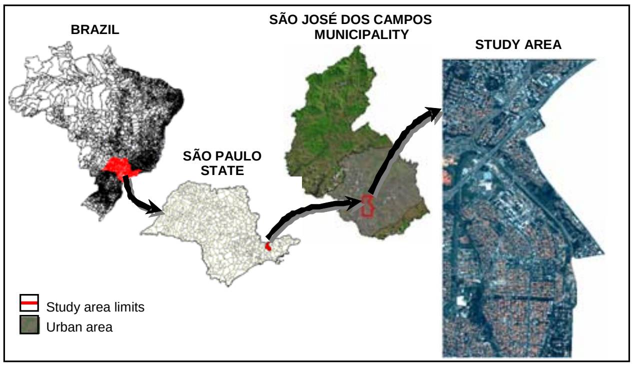


Figure 1: Study Area

A Quickbird scene (Ortho-ready Standard) was used in this study: a panchromatic image, 0,60 m spatial resolution and a multispectral image, 2,40 m spatial resolution, with 4 bands (blue, green, red and infrared). The images from May 17th 2004, have an off-nadir incidence angle of 7,0º and a radiometric resolution of 16 bits. we also used one street block layer related to the study area (in shape file format). This layer, issued by the planning department of the local government (PMJC, 2003), provides the elaboration of spatial rules to restrict the occurrence of some land cover classes.

We used the eCognition™ 4.1 software (DEFINIENS, 2003) for land cover classification and WEKA 3.5.6 software to generate the decision tree in one of the experiments. WEKA (Waikato Environment for Knowledge Analysis) is an open source software that contains a collection of machine learning algorithms for data mining tasks (available on http://www.cs.waikato.ac.nz/ml/weka/).

3. METHODOLOGY

Two experiments were carried out: 1) the images were classified using the e-Cognition™ 4.1 for the entire procedure and (2) the decision tree was generated using the C4.5 algorithm and afterwards the decision tree was converted to semantic network in the e-Cognition environment to classify the images. The classification results were compared to corroborate the fitness of C4.5 for structure knowledge in intra-urban land cover classification.

3.1 Preprocessing

Fusion methods based on PC transformation have shown good results in urban high resolution image analysis (Hoffmann, 2001). Therefore, a synthetic color image with better spatial resolution (0,6m) than that of the original multispectral images, produced by PC based fusion method, was used in the classification processes conducted in this work.

On the other hand, the HIS transformation was applied to generate other information layers: Intensity, Hue and Saturation components.

3.2 Class characterization

In this stage we selected the classes of interest based on the visual interpretation of the fused image. Different types of objects such as roofs, pavements and vegetation were identified. The attributes used to characterize the classes were organized in a key interpretation with information of color, form, context, texture, size and location. This key interpretation helps the specialist to create the semantic network for the object oriented image analysis.

The selected classes were: Objects of High Brightness (Light Concrete/Asbestos Roofs and very bright metallic roof), Trees, Grass, Shadow, Asphalt, Bare Soil, Ceramic Roofs, Dark Concrete/Asbestos Roofs, Medium Tone Concrete/Asbestos Roofs, Metallic Roofs, and Swimming-pools.

3.3 Experiment I – object oriented image analysis

This experiment is performed in two phases: multiresolution segmentation and semantic network building. In the first phase, given the number of segmentation levels, the parameters for each level such as scale, form and compactness are defined. In the second phase, a set of attributes is selected and a semantic network is created. Subsequently, the image objects are classified using this semantic network.

As our main focus is the structure knowledge, we will detail only the processes of the second phase. Firstly, the specialist provides a set of training samples (267 polygons) to supply the characterization of classes. In the semantic network, the easiest distinctive classes such as Vegetation (high NDVI values), Shadow (low Bright) and Objects of High Brightness (high bright) are put on its top.

The other classes could be analyzed using the *feature space optimization* tool available in the e-Cognition™ software. The *feature space optimization* uses the nearest neighbor classification algorithm to calculate the separability. From a set of attributes this tool points out the best combination that produces the largest separability among the classes. Unfortunately, this tool was not used in our experiments because the number of attributes was so great that the computer could not process the data. To overcome this problem, histograms of each pair of classes were compared (Figure 2) and the class with the shortest overlap was chosen. A set of 133 attributes was tested.

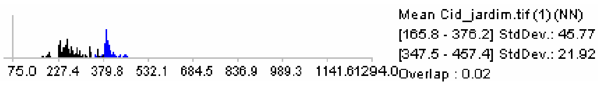


Figure 2: Histogram of the attribute "Mean" for two classes: Shadow (black bars) and Dark Concrete/Asbestos Roofs (blue bars).

As mentioned before, the semantic network construction is time-consuming because it depends mainly on the specialist knowledge. In our experiment, this stage took more than 600 hours for a test area of 12 km^2.

3.4 Experiment II- C4. 5 Classification

The C4.5 algorithm works in the following way. Each node of the decision tree matches an attribute and each arc matches a value range of that attribute. The expected attribute value is defined by the path from the root to each leaf.. The most representative attribute is associated to each node. The entropy is calculated to assess how informative a node is. The larger the entropy the more information is necessary to characterize the data. The goal is to associate the attribute which minimizes the data entropy to a node (Silva, 2007).

After associating the attribute to a node, the decision tree algorithm defines a threshold value for each arc. The threshold is computed by nearest neighbor algorithm. Firstly, the algorithm calculates the Euclidian distance from the training samples to an instance of the data. The instance will be assigned to the class that is closest to it in the space of attributes (Written, Frank, 1999).

To produce the shortest tree, the C4.5 algorithm removes unnecessary nodes through the pruning procedure. Thus, generating more generalized classes (Written, Frank, 1999) and a shorter tree. Also, the number of instances in each leaf also controls the size of the tree. The lower the number of instances the more precise the classification for the training set is. However, the algorithm can produce a complex tree with more inaccurate results for another set of samples. An alternative is to test various threshold values and to observe when the tree stabilizes. In other words, one selects a threshold value that does not produce significant changes in the accuracy classification.

In order to preserve the integrity of the evaluation, we used the same image objects for both experiments. We controlled the decision tree size through the minimum number of instances in each leaf. For each model, we calculated the kappa coefficient using the same 374 evaluation sample set of Experiment I.

Various tree models were tested. The best result was the tree with 5 instances per leaf, kappa coefficient value of 0.5052, and 14 nodes. Thereafter, this decision tree was converted to a semantic network in e-Cognition™ and used to classify the image objects.

4. RESULTS AND DISCUSSION

The classification results of the experiments were compared acording to the following criteria: (1) time for building the semantic network; (2) complexity of the semantic networks; (3) number of select attributes in each experiment, and (4) classification accuracy.

In relation to the time spent on the semantic network construction, the C4.5 model was far superior to the model created by the analyst in Experiment I. Only 40 hours was spent to produce the tree and classify the image objects whereas. more than 600 hours was spent on Experiment I.

Figures 3 and 3 show the semantic networks for both experiments. The classes represented in bold are classes of interest while the others are abstract ones. Abstract classes are superclasses that group classes with similar attribute values. As the object oriented analysis works with attribute heritage, the use of abstract classes increase the number of classification rules for each class of interest. Thus, the higher the number of abstract classes the more complex the network is. In figures 2 and 3, we can observe that the semantic network has 19 and 12 abstract classes in experiments I and II, respectively. We conclude that the second network is simpler than the first one.

Table 1 shows the selected attributes for each experiment. We observe that the number of attributes is the same for both experiments. Both experiments did not use geometric attributes but included attributes calculated from HSI components. We can also notice that the same three attributes were selected for both classifications: Ratio band 3/ band 1; Ratio band 2 and Mean diff. to neighbors band 1.

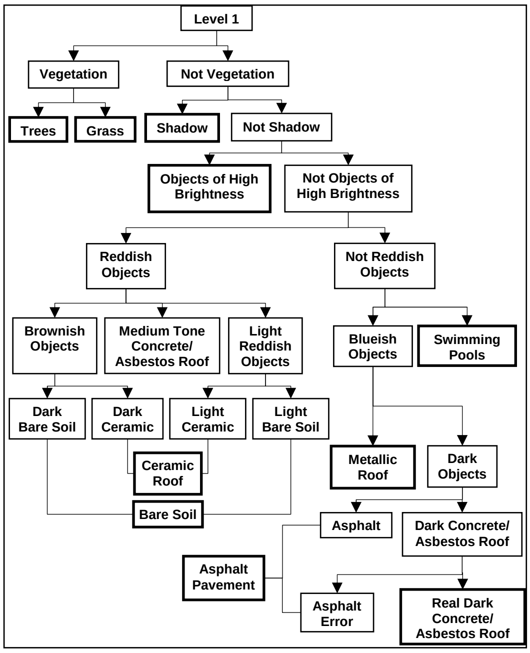


Figure 3 – Experiment I: semantic network.

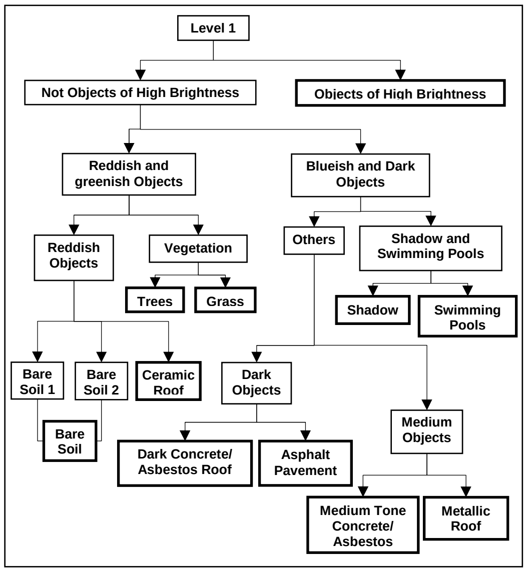


Figure 4 - Experiment II: semantic network.

Experiment I	Experiment II
Brightness	Standard Deviation Saturation channel
Mean of hue channel	Max pixel value Intensity channel
Mean of 3 band	Ratio band 3
Belongs to the super-object Block	Ratio to scene band 2
Max pixel value band 1	Stddev. Diff. to Super-object band 4
NDVI	Ratio band 1
Ratio band 3/ band 1	Ratio band 3/ band 1
Ratio band 2	Ratio band 2
Mean diff. to neighbors band 1	Mean diff. to neighbors band 1

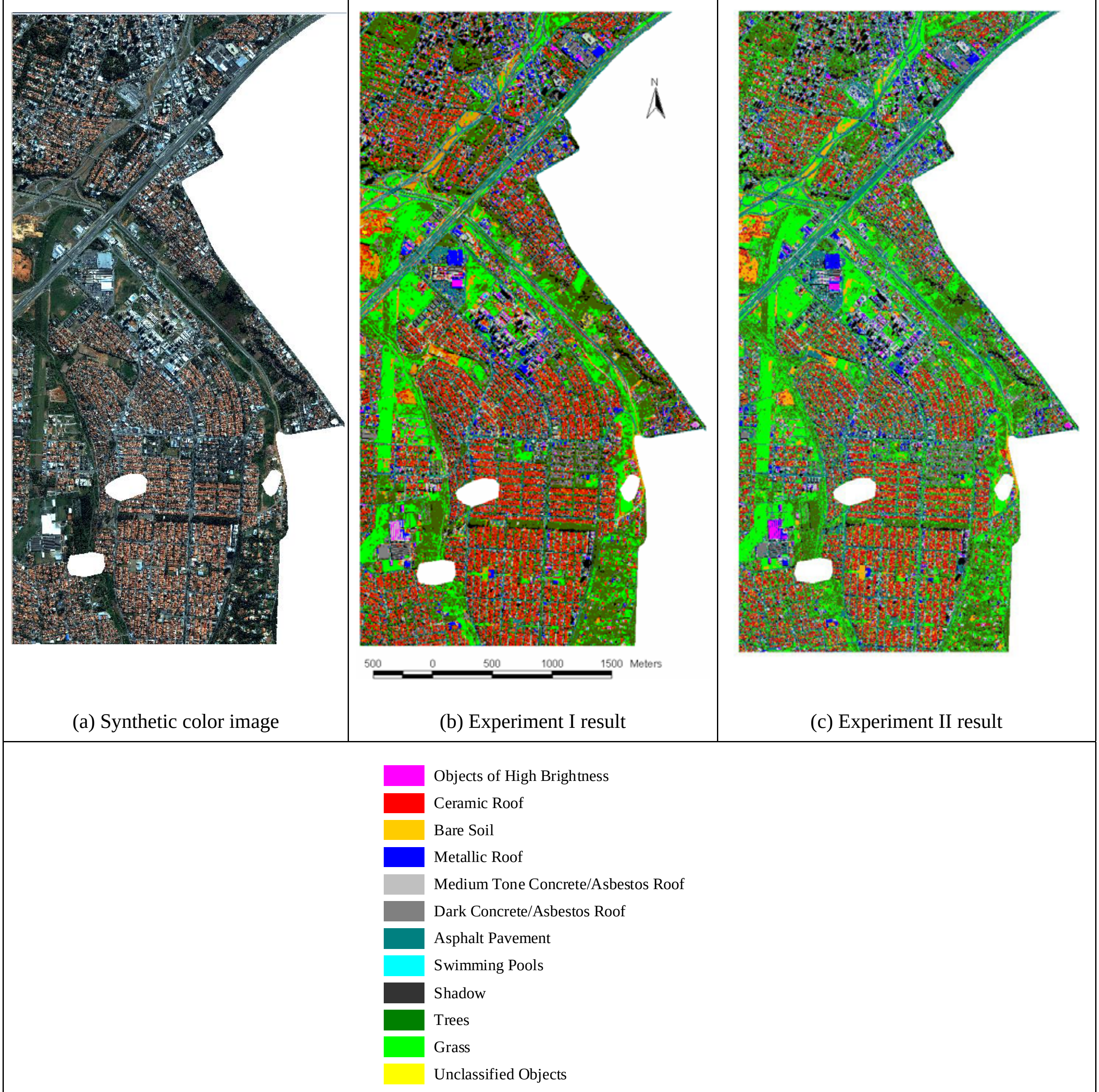
Table 1 - Selected attributes for each experiment.

In relation to thematic accuracy, the results were very similar. The kappa coefficient values were 0.5283 and 0.5052, respectively, for experiment I and II. According to Landis and koch (1977) both experiments produced good results.

Figure 5 shows the classification results. Visually, the results are very similar, as concluded before through quantitative evaluation.

5. CONCLUSION

This work aimed to evaluate how the C4.5 algorithm could help semantic network construction and attribute selection processes. Two experiments for intra-urban land cover classification were conducted: one using the e-Cognition system and the other using the C4.5 algorithm. The evaluation showed that C4.5 algorithm was very efficient in relation to the time spent on the semantic network construction and attribute selection tasks although the classification results were very similar. Moreover, the C4.5 semantic network was simpler than the one generated by the specialist in e-Cognition™. Finally, we can say that C4.5 algorithm integrated with other image analysis tools such as multiresolution segmentation and classification methods has great potential in high resolution image analysis for urban applications.



(a) Synthetic color image

(b) Experiment I result

(c) Experiment II result

Figure 5 – Classification results.

ACKNOWLEDGEMENTS

The authors would like to thank Imagem Soluções inteligência Geográfica for the Quickbird images as well as Brazilian National Council Scientific Research (CNPQ) for Carolina Moutinho Duque de Pinho´s PHD scholarship.

REFERENCES

Aksoy, S., et al. Interactive Training of Advanced Classifiers for Mining Remote Sensing Image Archives. In: ACM International Conference on Knowledge Discovery and Data Mining. 2004. Seattle, WA: ACM.

Centeno, J.; Miqueles, A. Extraction of buildings in Brazilian urban environments using high resolution remote sensing imagery and laser scanner data. In: ISPRS Congress, 20., 2004, Istanbul, Turkey. *Proceedings...* Istanbul: ISPRS, 2004. p. 589 - 692.

Datcu, M., et al., Information Mining in Remote Sensing Image Archives - Part A: System Concepts. *IEEE Trans. on Geoscience and Remote Sensing*, 2003. 41(2923--2936).

Definiens, *eCognition: User Guide 3*, 2003. 480 p. http://www.definiens-imagigng.com/down/ecognition. (accessed 15 out. 2003)

Herold, M., Liu X. & Clarke, K. C., Spatial metrics and image texture for mapping urban land use. *Photogrammetric Engineering and Remote Sensing*, v. 69, n. 9, p. 991-1001. Sept. 2003.

Hofmann, P; Detecting informal settlements from IKONOS image data using methods of object oriented image analysis - an example from Cape Town (South Africa). In: International Symposium Remote Sensing of Urban Areas, 2., 2001, Regensburg, Germany. *Proceedings…* Regensburg, Germany: Institut für Geographie an der Universität Regensburg, 2001(a). p. 107 - 118.

Landis, J. R.; Koch, G. G. The measurement of observer agreement for categorical data. *Biometrics*, v. 33, n. 1, p. 159-174. 1977.

Pinho, C.M., Kux, H. Dados do Quickbird para subsidiar o planejamento urbano: uma proposta metodológica, município de São José dos Campos, SP, Brasil. In: Simposio Latinoamericano sobre Percepción Remota y Sistemas de Información Espacial, 11., 2004, Santiago. *Proceedings…*Santiago: SELPER, 2004. CD-ROM.

Pinho, C. M. D. Análise orientada a objetos de imagens de alta resolução espacial. São José dos Campos. 2005. Master´s Thesis - National Institute for Space Research.

PMSJC - Prefeitura Municipal de São José dos Campos. 2003. Cidade Viva. [CD-ROM], São José dos Campos: PMSJC, 2003.

Thomas, N.; Hendrix, C.; Congalton, R. G. A comparison of urban mapping methods using high-resolution digital imagery. *Photogrammetric Engineering & Remote Sensing*, v. 69, n. 9, p. 963-972, Feb. 2003

Rego, F. G.; Koch, B. Automatic Classification of land cover with high resolution data of Rio de Janeiro city Brazil: comparison between pixel and object classification. In: International Sympsosium Remote Sensing of Urban Areas, 4., 2003, Regensburg. *Proceedings…*Regensburg: University of Regensburg , 2003. p. 153 – 157.

SILVA, M. P. S. Mineração de padrões de mudanças em imagens de sensoriamento remoto. 2006-02-03. 128 p. (INPE-14429-TDI/1130). PHD´s Thesis - National Institute for Space Research.

SILVA, M. P. S. ; CÂMARA, Gilberto ; ESCADA, Maria Isabel Sobral ; SOUZA, Ricardo Cartaxo Modesto de . Remote Sensing Image Mining: Detecting Agents of Land Use Change in Tropical Forest Areas. *International Journal of Remote Sensing*, 2007. in press.

Witten, I. H.; Frank, H. *Data mining*: practical machine learning tools and techniques with java implementations. San Francisco: Morgan Kaufmann, 1999.