On combining unsupervised classification and ontology knowledge

Germain Forestier, Cédric Wemmert and Pierre Gançarski

LSIIT - CNRS - University Louis Pasteur - UMR 7005
Pôle API, Bd Sébastien Brant - 67412 Illkirch, France
Email: {forestier,wemmert,gancarski}@lsiit.u-strasbg.fr

***Abstract*—This paper presents a way to combine knowledge obtained from a clustering algorithm and from an ontology. Using the both sources of information allows to improve the results of the knowledge discovery process. The basic property of clustering algorithms, which is to group similar objects, is the key of this approach. We use it to extend the knowledge given by an ontology. Indeed, this knowledge can be partial or not enough accurate, and clustering can then be used to fill this lack of information. We also present results and validation in the field of remote sensing image interpretation.**

***Index Terms*—Image classification, Pattern recognition, Clustering methods, Knowledge based systems**

I. Introduction

Unsupervised classification, also called clustering, is the process of organizing a set of data objects into homogeneous groups without availability of training samples or prior knowledge about the data. These groups, also called clusters, are collections of objects which are similar, but are dissimilar to the objects belonging to the other clusters. A huge number of clustering methods have been developed and several ones have been used for remote sensing image classification [1].

On the other hand, with the increase of spectral and spatial resolutions, knowledge based systems [1] have been more attentively investigated during the last few years, to improve image interpretation. Indeed, the so called *object oriented* [2] approach provides a new paradigm of reasoning by focusing on the objects present within an image, and not only on the pixels. In this field, a growing interest has been recently given to ontology-based representations [3], where the knowledge is stored as concepts with relations between them. This representation provides an easy and knowledgeable model, which stores information given by the domain expert about the geographical objects potentially present in an image (e.g. *house*, *road*, *vegetation*...).

In this paper we propose to combine these two sources of knowledge. On the one hand, an ontology of geographical objects which stores the knowledge about objects present in an image. On the other hand, a clustering algorithm which regroups similar objects together. The aim of our approach is to couple these two sources of knowledge (ontology and clustering) in order to fill the gap of knowledge produced by the lack of information stored in an ontology.

The first step of our approach is to identify the objects (i.e. the regions issued from a segmentation) which are recognized by the ontology. An unsupervised classification algorithm is then used on all objects, recognized by the ontology or not, to create different clusters of objects. Finally, in each cluster, the information obtained by the ontology on some objects of the cluster will be used to give information to the other objects of the clusters (not recognized by the ontology).

The paper is organised as follow. We first present the ontology and the knowledge-based object identification and how this knowledge is combined with a clustering algorithm. Then, we present an evaluation of the method for VHR remote sensing images interpretation. Finally, we conclude on perspectives of improvement of the system.

II. Description of the method

A. Description of the ontology

The regions obtained after an image segmentation associated to their features, are the input of the ontology-based object recognition. The method consists in matching each region with the concepts of an ontology. We have defined a matching measure and a traversing method of the ontology [3]. The proposed matching method is a feature-oriented approach. It verifies the validity of each feature values of one region according to the properties and the constraints defined in the concepts. The measure is composed of a local component (dealing with the inner properties of the concept) and a global component (evaluating the pertinence in the hierarchy of concepts). The local similarity measure $Sim(\mathcal{R},\mathcal{K})$ compares the features $\{v_i\}$ of a region $\mathcal{R}$ with the specific attributes of a concept $\mathcal{K}$. α_i is the weight of attribute a_i, expressing the role of a_i to recognize $\mathcal{K}$. $Valid$ evaluates the validity of an extracted feature v_i and the bounds of the accepted values of an attribute ($Valid(v_i, a_i)$=1 if v_i satisfies a_i). The matching score $Score(\mathcal{R},\mathcal{K})$ ($\in [0; 1]$) evaluates the pertinence of the matching between a region $\mathcal{R}$ and a concept $\mathcal{K}$ in the hierarchy of concepts. The matching score is a linear combination of local similarity measures obtained with the concepts $\mathcal{K}_j$ of the path, starting from the root of the ontology and ending at the studied concept ($\mathcal{K}_m$=$\mathcal{K}$). The local similarities are propagated by inheritance to more specific concepts. In this computation, we integrated a specialization coefficient based on the depth β_j of the concepts. In this way, the measure favours the specialization of the concepts, considering that all additional information give a new semantic.

$$Sim(\mathcal{R},\mathcal{K}) = \frac{\sum_{i=1}^{n} \alpha_i Valid(v_i, a_i)}{\sum_{i=1}^{n} \alpha_i} \quad (1)$$

$$Score(\mathcal{R}, \mathcal{K}_m) = \frac{\sum_{j=1}^{m} \beta_j Sim(\mathcal{R}, \mathcal{K}_j)}{\sum_{j=1}^{m} \beta_j} \quad (2)$$

The matching score between a region and a concept being defined, it remains to traverse the ontology to find the best concept(s) for a region, according to its score. We developed a level-wise algorithm to traverse the ontology, using heuristics to reduce the search space. The main heuristic corresponds to the selection of the best concepts at each level in order to prune some branches, having an irrelevant starting concept (with a poor matching score value). This strategy is based on the fact that an internal concept has more general properties than its children. If a few of these properties (or none) are valid, its children will not be relevant.

Finally, a function $\mathcal{F}$ is defined to return the concept with the maximal score for a region $\mathcal{R}$. A threshold *ScoreMin* is defined ($\in [0;1]$) as the minimal score where a region will be considered as a member of the studied concept.

$$\mathcal{F}(\mathcal{R}) = \begin{cases} \mathcal{K}_m \text{ where } Score(\mathcal{R}, \mathcal{K}_m) = max\{Score(\mathcal{R}, \mathcal{K}_i)\} \\ \text{if } Score(\mathcal{R}, \mathcal{K}_m) > ScoreMin \\ \\ unknown \text{ else} \end{cases} \quad (3)$$

B. Combining the knowledge

In this section, we explain how the two sources of knowledge are combined. First, each region is matched with the ontology and obtain an assigned concept with the maximum score value, or (for the most part) the *unknown* concept. Then, all the regions (identified or not) are grouped using a clustering algorithm. The basic idea is to observe for each cluster, the *majority concept* which is the most represented concept within this cluster. After having identified this concept, we are able to give a semantic (i.e. to affect a concept) to the *unknown* objects, by affecting the *majority concept* to these objects. With this mechanism, we extend the knowledge of the ontology. Let us formalize this procedure:

Let $\mathfrak{C} = \{C_i\}_{i=1\dots n_c}$ be a clustering with n_c clusters C_i.

Let $\Re_i = \{\mathcal{K}_j\}_{j=1\dots n_k^i}$ be the set of concepts identified within the cluster C_i.

To each cluster C_i is assigned a set of couples $\mathcal{L}_i$, each composed of a concept $\mathcal{K}_j$ identified in C_i, and the sum $\mathcal{S}_j$ of the scores of the objects of $\mathcal{K}_j$ in C_i:

$$\mathcal{L}_i = \{(\mathcal{K}_j, \mathcal{S}_j), \mathcal{K}_j \in \Re_i \setminus \{unknown\}, 1 \leq j \leq n_k^i\} \ , \quad (4)$$

$$\mathcal{S}_j = \sum_{\mathcal{R} \in C_i \,|\, \mathcal{F}(\mathcal{R}) = \mathcal{K}_j} Score(\mathcal{R}, \mathcal{K}_j) \ . \quad (5)$$

The *majority concept* $\mathcal{K}_i^{max}$ of the cluster C_i is defined as the one having the best sum of scores within the cluster:

$$(\mathcal{K}_i^{max}, \mathcal{S}_i^{max}) | \mathcal{S}_i^{max} = max\{\mathcal{S}_j\}_{(\mathcal{K}_j, \mathcal{S}_j) \in \mathcal{L}_i} \ . \quad (6)$$

The method computes $\mathcal{L}_i$ for each cluster C_i and then assigns the concept $\mathcal{K}_i^{max}$ to the *unknown* regions of C_i.

Fig. 1. The segmented image (900×900 pixels, resolution: 0.7m per pixel).

III. Experiment and Evaluation

In this section, we present an experiment on a VHR remote sensing image. The image selected for the experiment is from Strasbourg (France) and has been taken by Quickbird sensors. This area is representative of the urban structure of Western cities and is characterized by many different objects (e.g. buildings, streets, water, vegetation) that exhibit a diverse range of spectral reflectance values. We used an *object-oriented* approach which consists in segmenting the image, and in using the produced regions as data objects, described by the different properties of the regions (spectral, shape).

A. Object-Oriented image analysis

The first step in *object-oriented* remote sensing image interpretation is to obtain a good segmentation. Many algorithms are available for remote sensing image segmentation [4]. In this paper we used a method presented in [5]. The idea of this method is to apply a watershed transform on a fuzzy classification of the image to obtain the segmentation. The Fig. 1 presents the result of the segmentation where borders of each region are highlighted in white.

B. Geographical ontology

The used ontology (see Fig. 2) is composed of 91 concepts, 20 attributes and 66 final concepts. It has been designed in collaboration with geographer experts. This geographical ontology defines a set of concepts (buildings, water,...) and their relations. Each concept is defined by some low-level descriptors associated to intervals of accepted values (spectral, shape). After an image segmentation, the regions are fed into a concept selection module using the ontology to select the most plausible concepts (i.e. classes). For this task, low-level

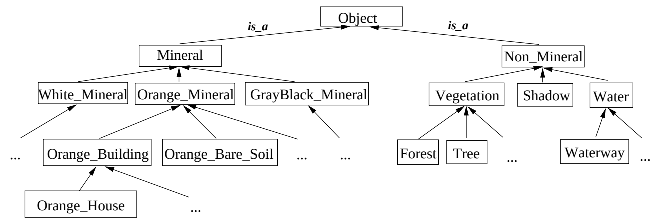

Fig. 2. Excerpt of the ontology.

descriptor values of the regions are computed and compared with the descriptors of the concepts from the ontology.

C. Collaborative clustering method

For the clustering part, we used a multi-strategical clustering method. This system integrates different algorithms of unsupervised classification and makes them collaborate to produce a unified result. A description of the method and its application to object-oriented image analysis can be found in [6]. We used this multi-strategical unsupervised clustering method on the regions obtained by the segmentation, characterized by spectral and spatial attributes.

D. Evaluation

To evaluate the pertinence of our approach we made several experiments. The Fig. 3 presents thumbnails representing the different concepts identified within the image: (a), (b), (c) and (d) show the regions identified only using the ontology, and (f), (g), (h) and (i), those identified after the assignment of the majority concept to *unknown* regions. (e) and (j) represent the *unidentified* part of the image. The Table I presents different evaluations for each of these thumbnails: the percentage covered by the concept on the image, the number of regions of this concept and the number of pixels covered by the regions of this concept.

As we expected, we succeed to increase the number of recognized regions and to decrease the number of unknown regions from 10991 *unknown* regions representing 65.72% of the image to 9546 *unknown* regions representing 36.61% of the image which is a significant improvement. Although, to validate the efficiency of our approach, we evaluated the classification accuracy obtained before and after the identification of the *unknown* regions. Indeed, we had to be sure of the legitimacy of the semantic given to the *unknown* regions. We used ground truth provided by manual labeling to compute for each class (i.e. concept) three known indexes to evaluate the quality of a classification, namely: *accuracy*, *recall* and *f-measure* (which is a tradeoff between accuracy and recall values).

The Table II presents the results of these three indexes for each thematic class. The first column gives the results for the ontology based approach and the second one the results for the proposed method (ontology and clustering). It appears that for all the 4 concepts (house, road, vegetation and water) the accuracy is stable, but the recall and f-measure hardly increase. This shows that the semantic assignment made using the clustering is relevant. The stability of the accuracy proves that the labeling of the *unknown* regions is correct. The increase of the recall indicates an augmentation of the number of identified regions.

			Onto.		Onto. + Clus.
house	% image		**5.67**		**9.66**
	#regions	(a)	87	(f)	158
	#pixels		45894		78272
road	% image		**6.65**		**11.04**
	#regions	(b)	176	(g)	381
	#pixels		53878		89406
vegetation	#pixels		**15.32**		**36.05**
	#regions	(c)	37	(h)	1160
	#pixels		124083		292009
water	% image		**6.64**		**6.64**
	#regions	(d)	9	(i)	9
	#pixels		53783		53783
unknown	% image		**65.72**		**36.61**
	#regions	(e)	10991	(j)	9546
	#pixels		532362		296530

TABLE I
EVALUATION OF AMOUNT OF THE IMAGE INTERPRETATION.

			Onto.		Onto. + Clus.
house	accuracy		0.934		0.910
	recall	(a)	**0.314**	(e)	**0.575**
	f-measure		0.470		0.705
road	accuracy		0.924		0.966
	recall	(b)	**0.507**	(f)	**0.695**
	f-measure		0.655		0.809
vegetation	accuracy		0.998		0.998
	recall	(c)	**0.471**	(g)	**0.973**
	f-measure		0.640		0.985
water	accuracy		1.0		1.0
	recall	(d)	**0.988**	(h)	**0.988**
	f-measure		0.994		0.994

TABLE II
EVALUATION OF THE ACCURACY OF THE IMAGE INTERPRETATION.

One can notice that we have not used a supervised approach with the identified regions as example, because all existing concepts are not present in the ontology. In our approach we keep the possibility to obtain unknown clusters, which contain objects not yet described in the ontology.

IV. Conclusion

In this paper we present a way to combine the knowledge from an ontology and from a clustering algorithm. The ontology gives information on few regions and a clustering algorithm is used to regroup all the regions in different clusters. The information obtained on some objects of each cluster is then used to send out the knowledge to the other members of the cluster. We present interesting results and an evaluation in the field of remote sensing image interpretation. We are now interested in the integration of the knowledge of the ontology directly within the collaborative clustering system to guide the classification process.

Acknowledgment

This work is a part of the FoDoMuST and ECOSGIL projects.

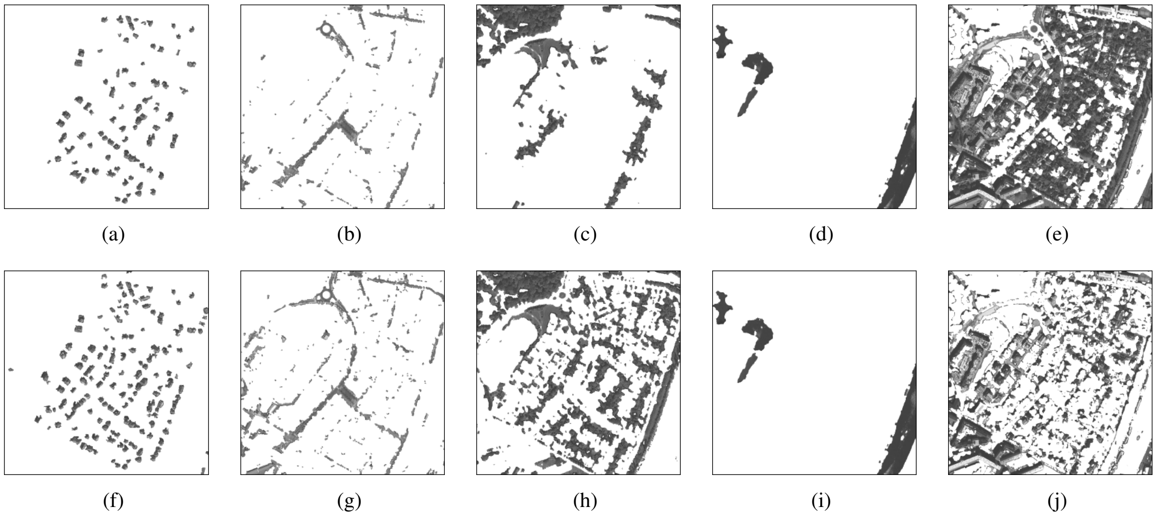

Fig. 3. Thumbnails of the QuickBird image representing the objects (*houses, roads, vegetation, water, unknown*) identified by the ontology (first row) and the objects identified by the combination with a clustering algorithm (second row).

References

[1] D. Lu and Q. Weng, "A survey of image classification methods and techniques for improving classification performance," *International Journal of Remote Sensing*, vol. 28, no. 5, pp. 823–870, 2007.

[2] L. Yongxue, L. Manchun, M. Liang, X. Feifei, and H. Shuo, "Review of remotely sensed imagery classification patterns based on object-oriented image analysis." *Chinese Geographical Science*, vol. 16, no. 3, pp. 282–288, 2006.

[3] N. Durand, S. Derivaux, G. Forestier, C. Wemmert, P. Gancarski, O. B. A., and Puissant, "Ontology-based object recognition for remote sensing image interpretation," in *IEEE International Conference on Tools with Artificial Intelligence*, vol. 1, 2007, pp. 472–479.

[4] A. P. Carleer, O. Debeir, and E. Wolff, "Assessement of very high spatial resolution satellite image segmentations," *Photogrammetric Engineering and Remote Sensing*, vol. 71, no. 11, pp. 1285–1294, 2005.

[5] D. S., L. S., W. C., and J. Korczak, "Watershed segmentation of remotely sensed images based on a supervised fuzzy pixel classification," in *Proc. the IEEE International Geosciences And Remote Sensing Symposium (IGARSS)*, 2006.

[6] G. Forestier, C. Wemmert, and P. Gancarski, *Supervised and Unsupervised Ensemble Methods and their Applications*, ser. Studies in Computational Intelligence, 2008, vol. 126/2008, ch. Collaborative Multi-Strategical Clustering for Object-Oriented Image Analysis, pp. 71–88.