New Metrics to Evaluate Pattern Recognition in Remote Sensing Images

Manel Kallel, Mohamed Naouai, and Yosr Slama

Faculty of Science of Tunis, University Tunis el Manar DSI 2092 Tunis Belvidaire-Tunisia,
manel.kallel@yahoo.fr, naouai@polytech.unice.fr,
yosr.slama@fst.rnu.tn

Abstract. The continuous development of pattern recognition approaches increases the need for evaluation tools to quantify algorithms performance and establish precise inter-algorithm comparison. So far, few performance evaluating metrics in pattern recognition algorithms are known in the literature, especially in remote sensing images. In this paper, four metrics are proposed for this purpose. The advantages and drawbacks of these metrics are first described, then some experimentation results are the presented in order to validate our contribution.

Keywords: Evaluating metrics, pattern recognition, performance evaluation, remote sensing.

1 Introduction

Pattern recognition (PR) covers a wide range of problems, and it is hard to find a unified view or approach. PR is used particularly in Engineering problems, such as character readers and waveform analysis, as well as to brain modeling in biology and psychology (brain modeling) [1].

The goal of a PR algorithm is to determine boundaries that separate forms of different classes and provide the best possible performance. PR consists of one of the following two tasks [2]: supervised classification (e.g. discriminant analysis) in which the input pattern is identified as a member of a predefined class, or unsupervised classification (e.g. clustering) where the pattern is assigned to a hitherto unknown class.

Performance evaluation of PR algorithms is more than an objective ; it is a necessity. In fact, there are two standard metrics, namely recall and precision, which are not reliable in some specific fields. We believe that these metrics do not accurately measure the different aspects of performance of PR algorithms especially in remote sensing images.

In this paper, we propose several metrics which can be used to determine how much the result of PR algorithms matches the ground truth. We attempted to propose these metrics as a result of a need for performance evaluation of PR algorithms in remote sensing images.

The remainder of the paper is organized as follows. In section 2, we detail our motivation. Section 3 is devoted to a brief state-of-the-art. In section 4, four proposed performance metrics are described. Section 5 first describes the experimental setup used to perform the evaluation and comparison of the metrics. Our experimental results and comments are then detailed. Finally, Section 6 provides a conclusion to this paper.

L. Alvarez et al. (Eds.): CIARP 2012, LNCS 7441, pp. 664–673, 2012.

2 Motivation (Remote Sensing Specificity)

Remote sensing images involve very specific forms and structures. Recognizing these objects and their spatial positions and evaluating the relationships between them is considered as a PR problem. However, their extraction is difficult and many works have been devoted to studying this topic.

Extracting objects from a satellite image is proved to be a hard problem because of its large variability. But, the difficulty degree depends on the types of existing scenes in images. In fact, there are two types of scenes i.e. city and rural ones.

A rural scene is different from a city scene. In the former, most of the area is farmland and most folk houses have similar appearance [3]. So rural visualization has a small amount of data, this makes recognition of objects easy. But the complexity of the urban landscape makes it difficult to produce efficient and reliable algorithms.

Several algorithms have been developed to extract the objects of these images for both types of scenes. The complexity of recognition and extraction vary, so we need reliable means to evaluate the performance of these algorithms and compare them.

So far, most of PR algorithms are evaluated through two metrics: precision and recall. Precision is the fraction of retrieved instances that are relevant, while recall is the ratio of relevant instances that are retrieved. However, counting the number of objects correctly recognized compared to those relevant or compared to those returned is not sufficient. We must also consider the exact location of these objects and their areas i.e. the number of pixels spanning these objects.

In fact, we consider five types of object recognition: correct recognition, over-recognition, under-recognition, misses, and noise. Over-recognition or multiple detections of a single object, results is not a correct recognition. Under-recognition, results is a subset of the correct recognition. A missed recognition is used when an algorithm fails to find an object which appears in the image (false negative). A noise is used when the algorithm assumes the existence of an object which is not in the image (false positive). Obviously, these measures may have various importance in different applications.

For these reasons, we propose metrics that will be most useful for evaluating performance for PR algorithms especially for remote sensing images.

3 Related Work

Performance evaluation is necessary for researchers to compare a new algorithm to those already existing and for users to choose an algorithm and adjust its settings depending on the problem to suit.

Several metrics have been proposed for evaluating PR algorithms. In object detection, Yi et al. [4] proposed a set of seven metrics for quantifying different aspects of a detection algorithm performance. As their names indicate, two of the metrics, i.e. Area-Based Recall for frame and Area-Based Precision for frame, are two variants of recall and precision, taking into account area of objects instead of their numbers. They are based on pixel count and they treat each pixel in the ground-truth as object/non-object and the output pixels as detected/non-detected.

Regarding evaluation of image segmentation with reference (ground-truth), Zhao and al. [5] suggested two precision measures to quantitatively evaluate the result segments of different algorithms. The region precision measures how many pixels are correctly classified and the boundary precision measures how close is the segment boundary to the real one. Yasnoff et al. [6] presented a new generalized quantitative error measure, based on comparison of both pixel class proportions and spatial distributions of ground truth and test segmentations.

Philipp-Foliguet and Guigues [7] proposed new criteria for evaluation image segmentation when no ground-truth is available. These criteria, based on an energy formalism, take into account both the complexity of the segmented image and the goodness-of-fit of an underlying model with the initial data. These evaluation criteria are thus multi-scale criteria. Various forms of energy formulation are experimentally compared.

4 Proposed Metrics

Our proposed metrics are described in the following subsections with their advantages and drawbacks. All the metrics values range from zero to one where one means every object is correctly recognized (perfect).

The first metric is general. It can be used to evaluate PR algorithms for any type of images including remote sensing images. It is inspired by the two metrics recall and precision. The second one is based on the area covered by the objects. The third one measures how well the ground truth is close to recognition result. The fourth has the same principle as the previous, except that the superposition is done object by object.

4.1 Object Correspondence (OC)

This metric is an object-count-based metric. It is inspired from both recall and precision metrics and merges them. It can be used for any type of image, not only remote sensing ones.

Let O_G be the set of relevant objects i.e. those of the ground-truth and O_R be the set of retrieved objects i.e. those of the result.

Therefore, recall and precision can be written in these forms:

$$recall = \frac{\mathrm{Card}(O_G \cap O_R)}{\mathrm{Card}(O_G)}$$

$$precision = \frac{\mathrm{Card}(O_G \cap O_R)}{\mathrm{Card}(O_R)}$$

We define the Correspondence Object metric CO as the ratio of relevant objects retrieved with the total of retrieved objects and relevant ones:

$$CO = \frac{\mathrm{Card}(O_G \cap O_R)}{\mathrm{Card}(O_G \cup O_R)}$$

This metric provides a single significant measure to compare different algorithms using the objects number criteria. In fact, if we want to compare the performance of two PR

algorithms, the two metrics recall and precision by returning two distinct values may not be enough significant, especially if the first one gives the best recall whereas the second gives the best precision as shown in Fig. 1. Therefore, it seems more interesting to combine the two metrics into a single one.

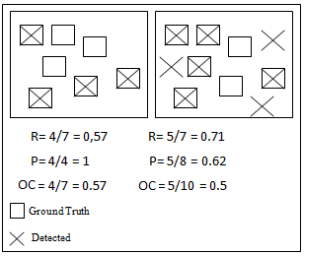

Fig. 1. Example illustrating Correspondence Object Metric

Despite its usefulness, the drawback of this metric as well as of both recall and precision is that they do not take into account the accuracy of the recognized object location or area relatively to the ground truth.

An object is deemed relevant or not using a visual assessment i.e. the user compares the retrieved objects to those of the ground truth and evaluates the relevance with the human eye. An over or under-recognized object can be declared as a relevant object.

In remote sensing images, another drawback is clear for the counting of objects because they can be infinitely small and the number is obvious.

4.2 Global Area (GA)

This metric is a pixel-count-based metric that measures how much the result of the algorithm approaches the ground-truth in terms of global area covered by the objects recognized in the same class. Among these classes, we find building, road, vacant land, vegetation and water.

We thought of offering this metric in order to solve the problem of counting objects that seem difficult. Instead of counting the number of objects and comparing the result with the ground-truth, we thought to proceed by their surface i.e. the number of pixels covered by these objects.

Hence, the metric consists of comparing the pixel number of the ground-truth objects of the same class with the pixel number of recognized objects. The number of pixels presents area or spatial union of objects of the same class.

Let$U_G$and $U_R$be the spatial union of all objects of the same class in ground-truth and result respectively. We have:

$$GA = 1 - \frac{|\mathrm{Card}(U_G) - \mathrm{Card}(U_R)|}{U_G}$$

To compute this metric for many classes, we can simply sum the values of this metric for all the classes and divide the sum by the class number.

Despite the ability of computing this metric, it presents a major drawback. In fact, the area of ground-truth objects may be close to the result objects, when recognition is not really well done (see Fig. 2).

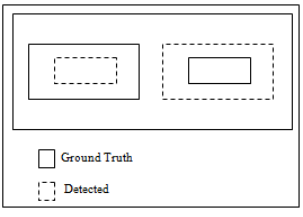

Fig. 2. Example illustrating GA metric limit

4.3 Superposed Area (SA)

This metric is a pixel-count-based metric that measures how much the result of the algorithm approaches the ground-truth.

The contribution of this metric compared to the previous is that it takes into account the detection error and not only the part well detected and this is defined by the union of the ground truth with the result of algorithm.

The objective of this metric is to determine the recognition rate for one class by taking into account the locations of objects. We have :

$$SA = \frac{\mathrm{Card}(U_G \cap U_R)}{\mathrm{Card}(U_G \cup U_R)}$$

The disadvantage of this metric is that the recognition of very close objects in one may produce a good result for the surface existing between these objects is very negligible see(Fig. 3).

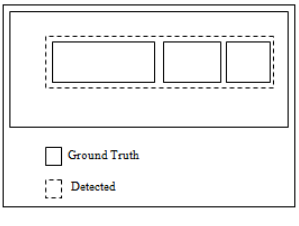

Fig. 3. Example illustrating SA metric limit

4.4 Superposed Per Object Area (SOA)

This metric is a pixel-count-based metric that measures how well the result of the algorithm approaches the ground-truth and assesses the spatial positions of the objects.

This metric considers first the number of pixels for each object and secondly its exact location. In this metric, the ground-truth and the algorithm output must be superposed to derive the recognized objects. A ground truth object is considered detected if a minimum proportion of its area is covered by the output object. A matching algorithm is used here to make correspondence between real and detected objects. It is based on maximizing the common area between correspondent objects. It should be noted that roads are treated as one object because they constitute a connected circuit. Thus, the matching algorithm principle is different from objects of other classes. In this case, many detected objects can be matched to the same ground truth object (road).

Let n be the total number of all the types of objects: correct recognition, over-recognition, under-recognition, missed, and noise, $U_{OG}(i)$ the spatial union of the object i in ground-truth and $U_{OR}(i)$ the spatial union of result object corresponding to i. We get :

$$SOA = \frac{\sum_{i=1}^{n} \mathrm{Card}(U_{OG}(i) \cap U_{OR}(i))}{\sum_{i=1}^{n} \mathrm{Card}(U_{OG}(i) \cup U_{OR}(i))}$$

5 Experimental Study

In order to provide a set of baseline results allowing a first evaluation of these metrics, we have considered the same experiments site on which we have applied three representative PR algorithms. The proposed metrics as well as recall and precision are then computed for each recognition result and each algorithm.

5.1 Experiment Site

For our experiment site, we consider a remote sensing image of the city of Strasbourg, France. Fig. 4 shows this image as well as the ground truth of the three classes it includes i.e. buildings, roads and vegetation. We have prepared these three images representing ground truth, especially for this work, as a first step in constructing a remote sensing image benchmark.

5.2 Experiment Algorithms

Three chosen PR algorithms to be evaluated in our experimental study are K-means [8], Particle swarm optimization [9] and Hierarchical Classification-based Region Growing [10]. These algorithms are able to detect buildings, roads and vegetations. We present in the following the description of each algorithm as well as the results we obtained.

K-Means. Algorithm K-means is the best known and most widely clustering algorithm due to the easiness of its implementation. It has been applied to our experiment image and has given the results shown in Fig. 5.

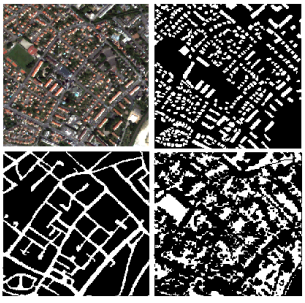

Fig. 4. Experiment site with buildings, roads and vegetation ground-truth

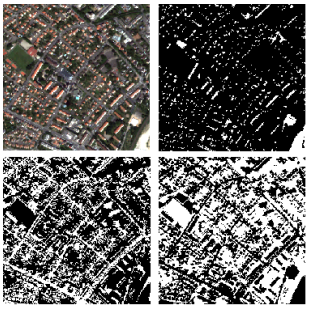

Fig. 5. K-means algorithm results

Particle Swam Optimization (PSO). The Particle Swarm Optimization algorithm belongs to the broad class of stochastic optimization algorithms that may be used to find optimal (or near optimal) solutions of numerical and qualitative problems. The recognition results of this algorithm is shown in Fig. 6.

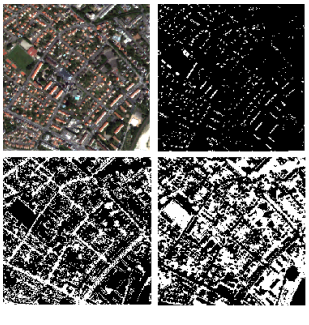

Fig. 6. PSO algorithm results

Hierarchical Classification-Based Region Growing (HCBRG). The algorithm is a hierarchical classification based on a region growing approach driven by expert knowledge represented in a hierarchical concept. A first classification will associate a

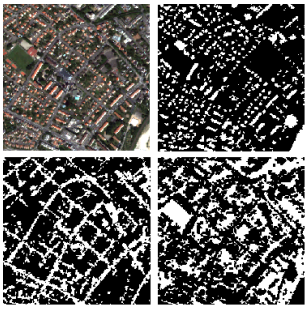

Fig. 7. HCBRG algorithm results

confidence score to each region in the image. This score will be used through an iterative step which allows interaction between segmentation and classification at each iteration. The region with the highest score will be taken as seeds in the growing step in each iteration and the approach allows the semantic growing based on the class of the seeds. This algorithm has given the results shown in Fig. 7.

5.3 Results

The performance evaluation of buildings, roads and vegetation recognition by the three chosen algorithms applied to the experiment image are depicted in Table 1.

We can notice from the above results that the values of each metric differ depending on both the PR algorithm used and the detected class. In fact, each algorithm has its own

Table 1. Proposed metrics values for three algorithms and three detection classes

Algorithms	Detected Classes		
	Buildings	Roads	Vegetation
K-means	GA = 0.38 SA = 0.19 SOA = 0.16 P = 0.41 R = 0.9	GA = 0.41 SA = 0.3 SOA = 0.3 P = 0.39 R = 0.93	GA = 0.36 SA = 0.55 SOA = 0.21 P = 0.14 R = 0.19
PSO	GA = 0.21 SA = 0.16 SOA = 0.14 P = 0.39 R = 0.85	GA = 0.65 SA = 0.33 SOA = 0.34 P = 0.41 R = 0.93	GA = 0.6 SA = 0.59 SOA = 0.28 P = 0.17 R = 0.35
HCBRG	GA = 0.8 SA = 0.41 SOA = 0.35 P = 0.7 R = 0.84	GA = 0.79 SA = 0.49 SOA = 0.49 P = 0.65 R = 0.92	GA = 0.89 SA = 0.9 SOA = 0.67 P = 0.71 R = 0.8

recognition ability and it depends on the nature of the objects to recognize. One can easily see that for the processed images, the first metric is sufficient to note that the last algorithm (HCBRG) is significantly better than the others for the three detected classes. This result was confirmed by the two other metrics giving more accurate comparisons. In addition, for the three metrics, K-means gives better (resp. worse) performances than PSO for buildings (resp. roads) recognition. However, we can, for example, remark that GA metric is not sufficient to accurately compare vegetation recognition by K-means and PSO algorithms. In fact, this metric leads us to think that PSO is much better than K-means (compare 0.6 to 0.36), whereas for SA and SOA metrics the two algorithms performance are quite close.

On the other hand, it is also noteworthy that the metrics values are in most cases decreasing in the order they are presented (i.e. GA then SA then SOA). This order is inversely proportional to the metrics accuracy as explained in their theoretical study. Therefore, these three metrics may be used according to the cascading priority criterion in our evaluation. Indeed, if the overall area of detected objects is the most important criterion, GA is the best to provide this information. However, if objects localization is also important, then SA can add more accuracy. SOA is the one that finally gives the most accurate information because it takes into account, in the same time, both the object sizes and their localization, and it also degrades for each missed or over detected object.

We finally note that only in the case of detecting roads, the last metric gives results very similar to the previous one. This is due to the different way of implementing connections between real objects and detected ones as explained in section 4.4.

6 Conclusion

Performance evaluation of image processing algorithms especially those of pattern recognition (PR), presents a major difficulty for researchers, because of the lack of performance evaluating metrics, particularly in the field of remote sensing. In this paper we proposed a set of metrics in order to provide specific and personalized performance evaluation of PR algorithms. The first metric, called Object Correspondence (OC), is based on object numbers and is proposed to combine the two metrics commonly used in pattern recognition i.e. Recall and Precision, in order to have only one value taking into account missed as well as over detected objects. The other three proposed metrics are rather specific to remote sensing images and principally based on the objects area. They have been introduced in an order proportional to their accuracy and inversely proportional to their respective complexities. The Global Area metric (GA) is based on a comparison of the global areas of ground truth objects and result ones. This metric is simply computed and is useful when one has a large number of small objects. However, it does not allow the evaluation of the good localization of detected objects. The Superposed Area metric (SA) measures how well the ground truth is close to recognition result by superposing their areas, This metric is usually most accurate than the previous one, but it does not take in account the corresponding objects. The last metric, called Superposition per Object Area (SOA) is finally proposed to allow the most accuracy level of performance evaluation. It takes into account localization and sizes of detected

objects as well as missed or over detected ones. We implemented the last three metrics on three significant algorithms: K-means, Particle Swarm Optimization and Hierarchical Classification-based Region Growing using as experiments site an RS image of the city of Strasbourg. The results obtained with these algorithms in buildings, roads and vegetation recognition were discussed in order to analyze the usefulness of each metric in performance evaluation. Thus, Our theoretical study of metrics has been validated. In addition, we deduced that the different metrics may be used in cascade according to chosen priorities of evaluation criterion. As a future work, it will be useful to extend our experimental study to other PR algorithms and other experiments sites. Besides, we intend to focus on another aspect in performance evaluation which is a standard benchmark construction for PR methods in remote sensing field. The experimental images of the benchmark must cover major challenges for PR such as light and shadow effects.

References

1. Fukunaga, K.: Introduction to statistical Pattern Recognition. Academic Press (1990)
2. Watanabe, S.: Pattern Recognition: Human and Mechanical. Wiley, New York (1985)
3. Li, D., Liu, Y., Chen, Y.: Computer and Computing Technologies in Agriculture IV. In: 4th IFIPTC 12 Conference, CCTA, China (2010)
4. Mariano, V.Y., et al.: Performance evaluation of object Detection Algorithms. Pattern Recognition 3, 965–969 (2002)
5. Zhao, Y., et al.: A benchmark for interactive image segmentation algorithms. In: Person-Oriented Vision (POV), pp. 33–38 (2011)
6. Yasnoff, W.A., Galbraith, W., Bacus, J.W.: Errormeasures for objective assessment of scene segmentation algorithms. AQC 1, 107–121 (1979)
7. Philipp-Foliguet, S., Guigues, L.: Evaluation de la segmentation d'images: état de l'art, nouveaux indices et comparaison. TS. Traitement du Signal 23(2), 109–124 (2006) ISSN 0765-0019
8. McQueen, J.: Some Methods for Classification and Analysis of Multivariate Observations. In: Proc. Fifth Berkeley Symp. Math. Statistics and Probability, pp. 281–297 (1967)
9. Kennedy, J., Eberhart, R.C.: Particle swarm optimization. In: Proceedings of the IEEE International Conference on Neural Networks IV, pp. 1942–1948. IEEE, Piscataway (1995)
10. Sellaouti, A., Hamouda, A., Deruyver, A., Wemmert, C.: Hierarchical Classification-Based Region Growing (HCBRG): A Collaborative Approach for Object Segmentation and Classification. In: Campilho, A., Kamel, M. (eds.) ICIAR 2012, Part I. LNCS, vol. 7324, pp. 51–60. Springer, Heidelberg (2012)