Comparing image segmentation algorithms for Content Based Image Retrieval Systems

Rami Albatal*, Philippe Mulhem,
Yves Chiaramella

MRIM, LIG, Rue de la Bibliothque,
Grenoble, 38041, France
**E-mail: Rami.Albatal@imag.fr*

Tat-Jun Chin

IPAL CNRS, I2R,
1 Fusionopolis Way,
138632, Singapore

This article discusses how to compare different image segmentation algorithms parameters in order to choose the most optimal algorithm parameters for a specific task(s) (e.g. feature extraction, spatial reasoning, topological analysis). Our method of comparison lets the user decide which segmentation algorithm/parameters are the most suitable for his system, this decision is obtained according to three indicators: the amount of relevant visual information as well as the noise in image regions, the average number of regions per images and the average number of regions per object.

Keywords: Image Segmentation Algorithms; Comparing Image Segmentation methods, Content Based Image Retrieval Systems

1. Introduction

This work takes place in the context of Content Based Image Retrieval (CBIR) systems. As shown in [11], such systems are composed of many different parts. Among them, the segmentation process is one of the first and important steps. Segmentation aims at separating the pixels of the images into consistent (according to any criterion) regions. In CBIR systems, it aims to facilitate the interpretation of images and scenes by identifying homogenous regions which can be further analyzed by the system by extracting region features and possibly their semantic meaning. CBIR approaches require good segmentation so that a large amount of relevant visual information is extracted. In CBIR systems, automatic segmentation algorithms try to extract significant regions in the picture, and regions from which we can extract relevant features. Ideally, a segmentation process should be able to remove regions that correspond to unimportant parts of an image. However, such algorithms are prone to errors due to the following reasons:

- Apart from the object-in-focus of the image (i.e., the important objects of the image), the resulting regions might contain unwanted background areas, for example in Fig. 1 where the region contains the ear of the sheep as well as grass. The visual information from the background is noise because it does not represent the object-in-focus, and because one object may occur on different backgrounds.
- Even though some regions do not include unwanted background areas (e.g. they are fully contained within the object-in-focus), they can be too small to allow reliable visual feature extraction, as presented in Fig. 2 where the plane is divided into many very small regions. Having a large number of small regions could have a bad effect or add more complexity on some task (like extracting the topological structure of objects). For instance, a large number of regions that describe an object may artificially complicate its visual structure, or even lead to a wrong description of this visual structure. For instance, the fuselage of the plane in Fig. 2 is split into many regions which do not really correspond to different parts of the plane.

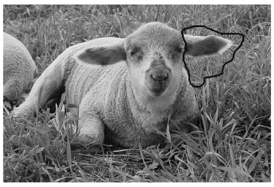

Fig. 1. Automatically segmented regions may include areas from different objects

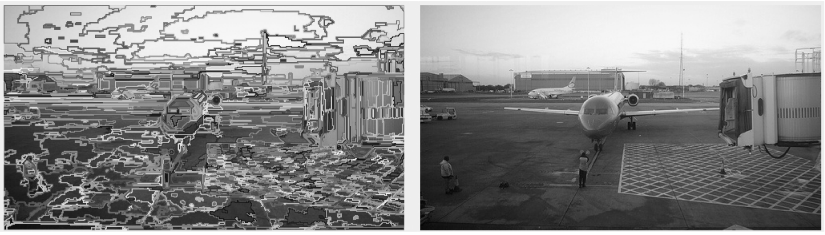

Fig. 2. An example of oversegmented image

For one specific CBIR system, it is important to choose a suitable automatic segmentation algorithm since it is a crucial step which affects the performance of all subsequent tasks. Most of the existing segmentation algorithms claim that they give good results [5,7,12], but they do not achieve the same level of performance for all images and corpora. Moreover, for one segmentation method, the problem of obtaining the best segmentation parameters remains. In this paper we investigate the issue of comparing image segmentation algorithms or sets of parameters of segmentation algorithms. We propose a measure for this purpose which is based on maximizing the amount of relevant visual information passed to the subsequent steps of the CBIR system, and we evaluate this measure on several segmentation schemes.

The rest of the paper is organized as follows: Related works are discussed in the section 2. In section 3, we explain the proposed measure. Our experiments are shown in section 4, Two segmentation methods will be explained and the parameters of each method will be tuned. Both methods will be compared using the VOC2008 segmented images collection[a]. Finally we conclude this paper in section 5.

2. Related works

Several evaluation methods have been proposed in the literature. Some methods tried to evaluate segmentation methods according to the visual feature used for segmentation, like color[1] or texture[4]. Other methods focus on comparing segmentation methods behaviors on specific types of images, like medical images [13,15], radar images[6], or noisy images[14]. More generic comparisons and evaluations methods are proposed in [2,3], and in [17,18] large spectrums of segmentation algorithms are scanned in order to compare them according to some criteria.

The work described here is close to [18] since we compare the result of an automatic image segmentation algorithm with an ideal manual segmentation by computing the overlapping area between these segmentations. However, our new metric is based on an F-measure indicator (well known in the field of Information Retrieval) and by taking into consideration two important indicators: the total number of regions in images and the average number of region per object. We assume that these elements are strongly related to the overall quality of a CBIR system.

[a] http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2008/

3. Evaluation method

In this work we try to remain generic, by proposing to a user (a developer of a CBIR system) some key criteria to help him choose the best segmentation algorithm (or the best parameters for a segmentation algorithm) for the given task expected from the CBIR. We do not propose to a user the best segmentation algorithm for an image collection without considering the user's interest and the later processing steps of the system. What we aim for with our method is to show to the user the effects for these algorithms or parameters with respect to what he needs.

As explained earlier, we use a ground truth for our evaluation. We obtain segmented regions of important objects (i.e. according to a specific need) from a subset of an image corpus. The manual segmentation is considered an "ideal" segmentation that the automatic method must seek to emulate. Then, we apply one or several automatic segmentations on the same subset and apply our evaluation method. We base our evaluation on the F-measure of the Information Retrieval (IR) domain. This measure has been used for many evaluation campaigns, like TREC[b], since the original paper describing it in [16].

In the context of IR, Recall and Precision values are defined in terms of a set of retrieved documents F and a set of relevant documents R:

$$Recall = \frac{|F \cap R|}{|R|} \qquad \text{and} \qquad Precision = \frac{|F \cap R|}{|F|}$$

The F-measure has the great advantage to combine both Recall and Precision values as their weighted harmonic mean. The traditional F-measure with equal weights of precision and recall is defined as:

$$F1 = \frac{2*(Precision*Recall)}{Precision+Recall}$$

We have to define now how to relate a segmentation result to the Recall and Precision values. In Fig. 3 we show a manually segmented region (circular), and automatic regions which overlap the manually segmented region. We consider the total area of automatically segmented regions as retrieved information, while the total area of manually segmented regions as relevant information. The remaining image areas that belong to the automatically segmented regions and that do not overlap the manually segmented region are considered non relevant retrieved information.

Let M be the set of all manually segmented regions m_j, and A the set of all automatically segmented regions a_k. All these regions are sets of pixels.

[b] http://trec.nist.gov/

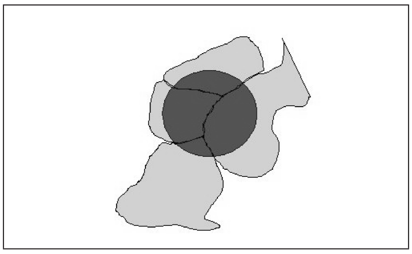

Fig. 3. Example manual (dark gray) and automatically segmented regions (light gray)

The set of regions of interest B (subset of A) contains the automatically segmented regions that overlap, at least with one pixel, a manually segmented region, as defined in:

$$B = \{a_k | a_k \in A \wedge \exists m_j \in M : m_j \cap a_k \neq \emptyset\}$$

Let us denote the set of overlapping areas between manually and automatically segmented regions:

$$O = \{m_j \cap a_k | a_k \in A \wedge \exists m_j \in M : m_j \cap a_k \neq \emptyset\}$$

The definitions of Recall and Precision are then:

$$Recall = \frac{\sum_{o_m \in O} |o_m|}{\sum_{m_j \in M} |m_j|} \quad \text{and} \quad Precision = \frac{\sum_{o_m \in O} |o_m|}{\sum_{a_k \in B} |a_k|}$$

Where $|X|$ denotes the number of pixels in a regions X.

According to the definitions above, and without any additional constraints, the sum of overlapped areas is always equal to the sum of areas of manually segmented regions. In fact, considering any overlapping region as a correct region for a automatically segmented region is unrealistic, because in a further step of the CBIR system we need to extract the features of these regions, so we need to ensure that a large ratio of such region actually corresponds to a manually segmented area. Regions with small overlap could have a negative effect on the learning stage due to an large background area. Here, if the overlap ratio of an automatic region is below a predefined overlapping threshold $T_{overlap}$ the whole region is excluded. We define then a set $O_{T_{overlap}}$ as:

$$O_{T_{overlap}} = \{m_j \cap a_k | a_k \in A \wedge \exists m_j \in M : \frac{|m_j \cap a_k|}{|a_k|} \geq T_{overlap}\}$$

By replacing O by $O_{T_{overlap}}$ in the Recall and Precision formulas above, we are able to describe the F measure of a segmentation algorithm, and therefore to compare these algorithms according to a reference.

4. Experiments

We evaluate the performance of a segmentation algorithm (or different versions of one segmentation algorithm) by computing the F-measure at a given overlapping threshold. After fixing the overlapping threshold we can change the parameters of any segmentation algorithm and compute the value of F-measure at the desired value of Precision or Recall. In this section we compare different parameters of a graph-based segmentation algorithm, and several regular grid segmentations.

4.1. *A Graph-Based image segmentation algorithm*

We choose here the graph-based image segmentation algorithm proposed by Felzenszwalb and Huttenlocher in [7] which has been used in [8–10]. An image is represented as an undirected graph G = (V,E), where each image pixel p_i has a corresponding vertex v_i in V. The edge set E is constructed by connecting pairs of pixels that are neighbors in an 8-connected meaning. An edge weight function is based on the absolute value of the intensity difference between the pixels connected by an edge.

A Gaussian blur filter is used to slightly smooth the image before computing the edge weights, in order to compensate for digitization artifacts. This Gaussian blur radius is the first parameter of this segmentation algorithm.

Pixels are grouped in regions according to the weight of connecting edges till that the region satisfies a condition related to a second parameter, (k). k sets the scale of observation: a large value for k causes a preference for larger regions. The third parameter, min_size, is a post processing parameter that enforces selected regions to have more pixels than min_size. If the region size is less than this number it will be merged to another region.
Here, we vary the value of the second parameter k since it is the most important one.

4.2. *A Grid-Based image segmentation algorithm*

Grid based segmentation is a simple way to segment images into regular rectangular blocks. This kind of segmentation does not rely on color or texture or any visual feature of the image, and images which have the same size will have the same segmentation. Fig. 4 shows an example of a grid based segmentation (with grid size of 16*16 pixels).

Here, We will apply different regular grid segmentations, we will vary the grid size (16*16 pixels, 32*32 pixels and 64*64 pixels).

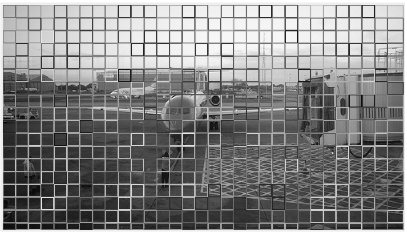

Fig. 4. Example of a grid based segmentation, grid size = 16*16 pixels

4.3. *Corpus*

The VOC2008 collection supplies 1023 manually segmented and annotated images. Fig. 5 shows an example of a manually segmented image from VOC2008 collection.

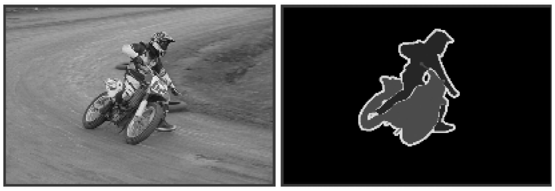

Fig. 5. Manually segmented image from VOC2008 collection

We randomly choose 300 manually segmented and annotated images from VOC2008 collection, chosen images most cover all the 20 object classes in VOC2008 collection. We then segment these images automatically using the graph-based segmentation algorithm and automatic regions which are deemed overlapping with the manual regions will inherit the label of the manual region the manual annotations into the resulting automatically segmented regions according to different degrees of overlapping thresholds between the automatic regions and the manual ones. The same process is also applied for the grid segmentation.

4.4. *Experiments results*

Fig. 6 shows the F-measure values at 11 different levels of overlapping thresholds ($T_{overlap}$ from 0% to 100% with a step of 10%).

As we can see in Fig. 6 the 16*16 pixels grid segmentation gives the highest F value at almost all of the overlapping thresholds. This result can

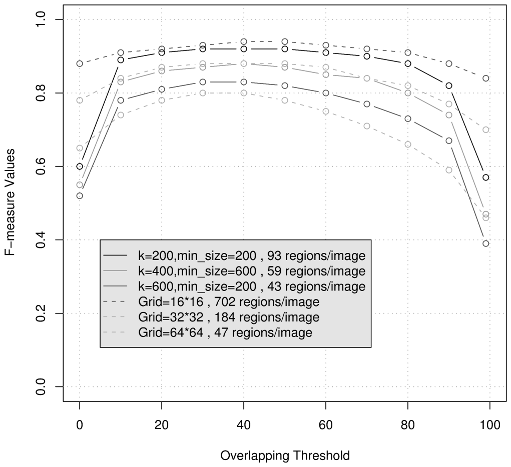

Fig. 6. F-measure values for different parameters values

due to the small size (256 pixels) of regions. The drawback of such segmentation however is the large number of resulting regions (703 regions per image), as well as the high average number of regions per object (e.g. 20 region for *aeroplan* objects, 16 for *horse* objects, and 18 for *sofa* objects). As mentioned in the introduction, this may complicate the following processes of the CBIR system. However if the segmentation method has to be simple to evaluate (like on a mobile device with small memory and/or small computing power), we see that this method is attractive. The second best result is achieved with the graph segmentation with k=200 and min_size=200, For instance, when we consider an overlap threshold of 80%, the difference of F-value compared to grid at 16*16 is only 3%. The great advantage of this segmentation is that the number of regions per image (93 regions per image) and the number of regions per object (e.g. 8 regions for *aeroplane* objects, 10 for *horse* objects, and 6 for *sofa* objects) are lower. Furthermore, the regions do not have the same shape and the same size which allows the analysis of the shapes of regions belonging to each objects as well as the shapes of objects.

This number of regions per object gives also to the user an idea about the complexity of the segmented objects: for instance, for the threshold 80% and the graph segmentation with k=200 and min_size=200, manually segmented region of *train* is split into 60 regions on average, while a more simple object like *dog* is split on average into 4 regions.

With the results provided, the user can fix an overlapping threshold, 80% for example, and then choose which segmentation to apply depending on the later stages of the CBIR he is building. Someone could prefer to have smaller number of regions and give less importance to the noise included in regions (as with the graph segmentation with k=600 and min_size=200). Another user could be more interested to have the maximum of relevant information in the regions and minimum noise regardless of the number of regions (as with the 16*16 pixel grid segmentation). A user could also be interested in studying the shapes of regions and in the same time he wants to have a large collection of regions with maximum of relevant information (in this case the graph segmentation with k=200 and min_size=200 is more suitable). So the F-measure as well as the average number of regions per image and per object can be used by a user to decide what segmentation to use.

5. Conclusion

In this paper we illustrated a simple method of comparing different segmentation algorithms or different parameter values for one segmentation algorithm. The proposed comparison does not favor one segmentation algorithm or a fixed values of parameters. It proposes to the user three indicators that help him to make a decision : the F-measure value which gives an idea about the noise and the relevant visual information in resulted regions, the average number of regions per image, and the average number of regions per object. The last two indicators give an indication about the complexity of potential post processing required on the regions.

In the future, we need to integrate more clearly the two elements of F-measure and the size of regions in a way that can clarify the choice to be made by a user. We will also validate our proposal on other collections besides VOC2008.

References

1. M. Borsotti, P. Campadelli, and R. Schettini. Quantitative evaluation of color image segmentation results. *Pattern Recogn. Lett.*, 19(8):741–747, 1998.
2. J. S. Cardoso and L. C. Real. Toward a generic evaluation of image segmentation. 14(11):1773–1782, November 2005.
3. S. Chabrier, B. Emile, C. Rosenberger, and H. Laurent. Unsupervised performance evaluation of image segmentation. *EURASIP J. Appl. Signal Process.*, 2006(1):217–217, 2006.

4. K. I. Chang, K. W. Bowyer, and M. Sivagurunath. Evaluation of texture segmentation algorithms. *Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on.*, 1:–299 Vol. 1, 1999.
5. Y. Deng and B. S.Manjunath. Unsupervised segmentation of color-texture regions in images and video. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI '01)*, 23(8):800–810, Aug 2001.
6. Y. Dong, B. Forster, and A. Milne. Evaluation of radar image segmentation by markov random field model with gaussian distribution and gamma distribution. *Geoscience and Remote Sensing Symposium Proceedings, 1998. IGARSS '98. 1998 IEEE International*, 3:1617–1619 vol.3, Jul 1998.
7. P. F. Felzenszwalb and D. P. Huttenlocher. Efficient graph-based image segmentation. *Int. J. Comput. Vision*, 2004.
8. D. Hoiem, A. A. Efros, and M. Hebert. Automatic photo pop-up. *ACM Trans. Graph.*, 24(3):577–584, 2005.
9. D. Hoiem, A. A. Efros, and M. Hebert. Geometric context from a single image. *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, 1:654–661 Vol. 1, Oct. 2005.
10. O. V. Kaick and G. Mori. Automatic classification of outdoor images by region matching. In *CRV '06: Proceedings of the The 3rd Canadian Conference on Computer and Robot Vision*, page 9, Washington, DC, USA, 2006. IEEE Computer Society.
11. M. S. Lew, N. Sebe, C. Djeraba, and R. Jain. Content-based multimedia information retrieval: State of the art and challenges. *ACM Trans. Multimedia Comput. Commun. Appl.*, 2(1):1–19, 2006.
12. T. Liu, J. R. Kender, R. Hjelsvold, and A. Pizan. A fast image segmentation algorithm for interactive video hotspot retrieval. *cbaivl*, 0:3, 2001.
13. G. Wagenknecht, H.-J Kaiser, T. Obladen, O. Sabri, and U. Buell. Simulation of 3d mri brain images for quantitative evaluation of image segmentation algorithms. In K. M. Hanson, editor, *SPIE , Medical Imaging 2000: Image Processing*, volume 3979 of *Presented at the Society of Photo-Optical Instrumentation Engineers (SPIE) Conference*, pages 1074–1085. Society of Photo-Optical Instrumentation Engineers, June 2000.
14. Y. Xu, V. Olman, and E. C. Uberbacher. A segmentation algorithm for noisy images: Design and evaluation. 19(13):1213–1224, November 1998.
15. J. Yang and S. C. Huang. Method for evaluation of different mri segmentation approaches. *Nuclear Science Symposium, 1998. Conference Record. 1998 IEEE*, 3:2053–2059 vol.3, 1998.
16. Y. Yang and X. Liu. A re-examination of text categorization methods. In *SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 42–49, New York, NY, USA, 1999. ACM.
17. Y. J. Zhang. A survey on evaluation methods for image segmentation. *Pattern recognition*, 29(8):1335–1346, 1996.
18. Y. J. Zhang. A review of recent evaluation methods for image segmentation. *Signal Processing and its Applications, Sixth International, Symposium on. 2001*, 1:148–151 vol.1, 2001.