Evaluation measures for segmentation

Guillaume Lemaître - Eng Wei Yong
Heriot-Watt University, Universitat de Girona, Université de Bourgogne
g.lemaitre58@gmail.com - engweiyong@gmail.com

I. Introduction

Segmentation is an essential process in image processing, medical imaging and machine vision. Segmentation evaluation research is dealt with the development of the tools and techniques to measure and compare the performance of the segmentation algorithms. Performance is highly depends on its applications. In some cases, computational efficiency and stability are essential. In another, the output is good if it resembles human perceptual grouping.

Many researchers eager in finding the new segmentation method but few are interested in developing the evaluation framework to compare different algorithms. Most of the researchers provide the method to evaluate their algorithm performance however a general standardized evaluation framework is lacking. Segmentation evaluation is still a new area which receiving less attention than the segmentation method itself.

II. Objective

Many segmentation methods have been studied and implemented in the image processing applications. It is essential to be able to evaluate and compare the segmentation methods. It is important to the application developer to choose the right tool for implementation. It is also essential for the researchers to evaluate and enhance new segmentation methods through formal comparison with the current methods.

III. Evaluation Criteria

It is hard to establish the evaluation measure for segmentation by considering various kind of the performance metrics required to meet the objective of the segmentation. However, generally segmentation performance is evaluated based on these three types of metrics: accuracy, precision and efficiency in order to avoid error in results.

- Accuracy: a measure of how well the segmentation output agrees with human perception.
- Efficiency: a measure of amount of time or effort required to perform segmentation.
- Precision: a measure of degree to which the same result would be produced over different segmentation sessions.

IV. Supervised evaluations

Supervised evaluations are used to find the quality of segmentation. These methods are called supervised because an absolute segmentation is used to compare with the segmented image obtained after performing segmentation algorithm . This absolute segmentation image is called "Ground-Truth". Supervised evaluations can be decomposed in three different methods:

- Evaluation Metrics Based
- Local and Global Consistency Error (LCE - GCE)
- Huang and Dom Evaluation Measure

A. Evaluation Metrics Based

In this section, we will present three different basic distances which give some information regarding the quality of segmentation algorithm. These distances are:

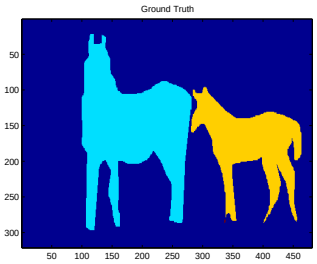
- Rand index
- Jaccard index
- Fowkles and Mallows index

In order to compute these different indices, a confusion matrix has to be computed:

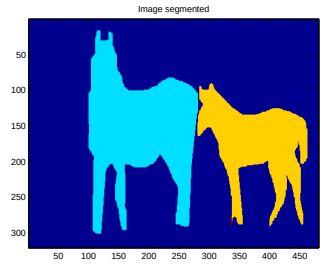
1) Confusion matrix: Assuming that a segmented image is composed of m segments and the Ground-Truth is composed of n segments. The confusion matrix linking both images is as shown on the table I.

M_{ij} represents the number of pixels belongs to the segment i of the segmented image S and j of the Ground-Truth GT.

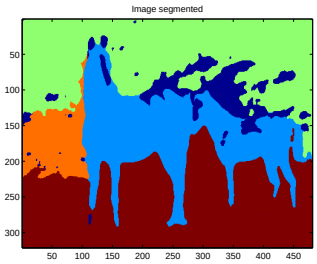
Table II represents the confusion matrix between the image 1(a) and 1(b) while the table III represents the confusion matrix between the image 1(a) and 1(c). In



(a) Ground Truth image



(b) First image segmented considering as perfect segmentation. This image was computed applying a dilatation on the Ground Truth



(c) Second image segmented obtained using a region growing algorithm

Figure 1. Set of image: Ground-Truth, perfect segmentation, region growing segmentation

	GT_1	GT_2	...	GT_n
S_1	M_{11}	M_{12}	...	M_{1n}
S_2	M_{21}	M_{22}	...	M_{2n}
...	...	...	...	...
S_m	M_{m1}	M_{m2}	...	M_{mn}

Table I
EXAMPLE OF CONFUSION MATRIX

	GT_1	GT_2	GT_3	GT_4
S_1	110107	0	0	0
S_2	1970	25447	0	0
S_3	2282	0	14566	0
S_4	20	0	0	9

Table II
CONFUSION MATRIX BETWEEN FIGURE 1(A) AND 1(B)

order to evaluate the future measure, we created a perfect segmentation, figure 1(b), which is a dilatation of the Ground-Truth. However, we compute a segmented image using region growing shown in figure 1(c). Figure 1(c) represents the absolute segmentation, Ground Truth, used to compute the comparison.

In order to compute the different distances, the different following number have to be computed:

- n_{11}: number of pixels which belongs to the same segment in the segmented image and the Ground-Truth.
- n_{10}: number of pixels which belongs to the same segment in the segmented image but not in the Ground-Truth.
- n_{01}: number of pixels which belongs to the same segment in the Ground-Truth but not in the segmented image.
- n_{00}: number of pixels which are in different region in the Ground-Truth and the segmented image.

	GT_1	GT_2	GT_3	GT_4
S_1	7780	1771	1249	0
S_2	1301	22027	11405	0
S_3	46681	276	42	0
S_4	9263	147	0	0
S_4	49354	1226	1870	9

Table III
CONFUSION MATRIX BETWEEN FIGURE 1(A) AND 1(C)

The following equations allow to compute the previous information:

$$n_{11} = \frac{1}{2}[\sum_{i=1}^{k}\sum_{j=i}^{l} M_{ij}^2 - n] \tag{1}$$

which represents the sum of the square of the diagonal of the confusion matrix minus the number of region in the Ground-Truth.

$$n_{10} = \frac{1}{2}[\sum_{i=1}^{k} |GT_i|^2 - \sum_{i=1}^{k}\sum_{j=i}^{l} M_{ij}^2] \tag{2}$$

which represents the difference between the number total of pixel each region of the Ground-Truth squared and the sum of the square of the diagonal of the confusion matrix.

$$n_{01} = \frac{1}{2}[\sum_{j=1}^{l} |S_j|^2 - \sum_{i=1}^{k}\sum_{j=i}^{l} M_{ij}^2] \tag{3}$$

which represents the difference between the number total of pixel each region of the segmented image squared and the sum of the square of the diagonal of the confusion matrix.

$$n_{00} = \frac{n(n-1)}{2} - n_{11} - n_{10} - n_{01} \tag{4}$$

which represents the number total of pixels minus the previous information computed.

	Rand	Jaccard	Fowkles
First segmentation	0.0476	0.0804	0.0415
Second segmentation	0.7951	0.9653	0.9278

Table IV
DIFFERENT VALUES GIVEN BY THE COMPUTATION OF THE DISTANCES

2) Rand Index: The first metric evaluation is named Rand index and allows to give the accuracy of the segmentation comparing the Ground-Truth and the segmented image. This measure represents the closeness of the Ground Truth and the segmented image. The formula allowing to compute the distance is the following:

$$R(GT,S) = 1 - \frac{n_{11}+n_{00}}{n(n-1)/2} \quad (5)$$

The distance tends to 0 if the segmented image is closed to the Ground Truth and tends to 1 when the difference between both images is important.

3) Jaccard Index: The second metric evaluation is named Jaccard index and allows to give the similarities of the segmentation comparing the Ground-Truth and the segmented image. The formula allowing to compute the distance is the following:

$$R(GT,S) = 1 - \frac{n_{11}}{n_{11}+n_{10}+n_{01}} \quad (6)$$

The distance tends to 0 if the segmented image is similar to the Ground Truth and tends to 1 when the difference between both images is important.

4) Fowkles and Mallows index: The third metric evaluation is named Jaccard index and allows to give the similarities of the segmentation comparing the Ground-Truth and the segmented image. The formula allowing to compute the distance is the following:

$$F(GT,S) = 1 - \sqrt{W_1(GT,S)W_2(GT,S)} \quad (7)$$

$$W_1(GT,S) = \sum_{i=1}^{k} \frac{n_i 1}{|GT_i|(|GT_i|-1)/2} \quad (8)$$

$$W_2(GT,S) = \sum_{j=1}^{l} \frac{n_1 1}{|S_j|(|S_j|-1)/2} \quad (9)$$

The distance tends to 0 if the segmented image is similar to the Ground Truth and tends to 1 when the difference between both images is important.

5) Results: Table IV presents the evaluation of the metric based evaluations presented in the previous parts. We can notice that the evaluation of the segmented image 1(b) gives a result near of 0. However, the evaluation of the segmented image 1(c) gives a result near of 1. Considering only these last results, the segmented image 1(c) should be considered as inaccurate. We can explain this phenomenon because we considered the Ground-Truth as absolute. Some other methods allow to moderate the weight of the Ground-Truth.

B. Local and Global Consistency Error (LCE - GCE)

In the previous part, the Ground-Truth was considered like absolute reference. However, due to the human perception, a Ground-Truth can change from the point of view of different specialists. The Local and Global Consistency Error (LCE - GCE) allows to evaluate the dissimilarities between the Ground-Truth and the segmented image and between the segmented image and the Ground-Truth.

1) Local refinement error: In order to compute the LCE and GCE, the local refinement error between clusters of the Ground-Truth and the segmented image and between clusters of the segmented image and the Ground-Truth. The error is defined as follow:

$$E(GT,S,p_i) = \frac{|R(GT,p_i) \backslash R(S,p_i)|}{|R(GT,p_i)|} \quad (10)$$

$$E(S,GT,p_i) = \frac{|R(S,p_i) \backslash R(GT,p_i)|}{|R(S,p_i)|} \quad (11)$$

2) Local Consistency Error - LCE: The LCE is defined as follow:

$$LCE = \frac{1}{n} \sum_{all\ p_i} \min E(GT,S,p_i), E(S,GT,p_i) \quad (12)$$

The distance tends to 0 if the segmented image is a good segmentation and tends to 1 if it is a bad segmentation.

3) Global Consistency Error - GCE: The GCE is defined as follow:

$$GCE = \frac{1}{n} \min \sum_{all\ p_i} E(GT,S,p_i), \sum_{all\ p_i} E(S,GT,p_i) \quad (13)$$

The distance tends to 0 if the segmented image is a good segmentation and tends to 1 if it is a bad segmentation.

4) Results: Table V presents the evaluation using LCE and GCE evaluation. We can notice that the evaluation of the segmented image 1(b) gives a result near of 0 while the result for the 1(c) is enough good because inferior to 0.2.

	LCE	GCE
First segmentation	0.0247	0.0493
Second segmentation	0.1171	0.1851

Table V
DIFFERENT VALUES GIVEN BY THE COMPUTATION USING LCE AND GCE

	Huang & Dom
First segmentation	0.9724
Second segmentation	0.7056

Table VI
DIFFERENT VALUES GIVEN BY THE COMPUTATION USING HUANG & DOM EVALUATION

C. Huang and Dom Evaluation Measure

The LCE and GCE evaluation is sensitive to degenerate case and to under or over classification. Huang and Dom evaluation measure allows to avoid this phenomenon and ignore refinement between two images. The Huang and Dom evaluation is defined as follow:

$$HD = 1 - \frac{D_H(GT \to S) + D_H(GT \to S)}{2A} \quad (14)$$

where A is the area of the image and D_H is the Hamming distance defined as follow:

$$D_H(GT \to S) = \sum_i \sum_{j \neq \max(i)} |GT_i \cap T_j| \quad (15)$$

$$D_H(S \to GT) = \sum_i \sum_{j \neq \max(i)} |S_i \cap GT_j| \quad (16)$$

The distance tends to 1 if the segmented image is a good segmentation and tends to 0 if it is a bad segmentation.

1) Results: Table VI presents the evaluation using LCE and GCE evaluation. We can notice that the evaluation of the segmented image 1(b) gives a result near of 1 while the result for the 1(c) is enough good because superior to 0.7.

D. Comparison

Results obtained with metric evaluations bad because these distances were not defined for segmentation evaluation initially and are strict. However, these methods can be combined. LCE and GCE tolerate refinement and are not strict as the metric evaluations. However, these evaluations are not performing well with over or under segmentation and are not working with degenerate cases. Contrary to LCE and GCE, Huang and Dom evaluation allows to perform good evaluation with over and under segmentation and degenerate cases. However, this method rejects refinement that can be a problem if we want to use this property.

V. UNSUPERVISED EVALUATIONS

Unsupervised evaluation measures don't use a reference ground truth. It is based on the fact that there are other properties that can be used to evaluate the segmentation performance without using ground truth. Human being is able to judge the output of the segmentation by purely observing it without the knowledge from ground truth. It can be judged intuitively that whether an image is prone to over-segmentation or under-segmentation by just observing. Besides, conclusion can be easily drawn from a segmentation output which is very jagged or symmetrically.

The objective of unsupervised evaluation is to measure the performance of a segmentation given only segmentation and its output. It is harder due to lesser information. Nevertheless, it is good that it can avoid ambiguous ground truth for complicated scenery. In addition, unsupervised evaluation measures can be used to choose the good values for parameters that affect a segmentation output automatically. The unsupervised evaluation might undergo training phase to learn what constitute a good segmentation algorithm.

Since unsupervised evaluation doesn't have a reference, it is purely based on a fundamental understanding of human perceptual grouping. Thus, the formulation of the objective of the segmentation problem is focused rather than the implementation method. That is to find the criteria which make good segmentation output and optimize the criteria. However, it is difficult to formulate an object for a segmentation problem. Thus, most of the unsupervised evaluation measures have limited success. There are still some successful measures as follow:

- Entropy-based evaluation: It is based on information theory.
- Visible colour distance based evaluation: It is based on perceived colour distances.

Entropy-based evaluation is discussed in the following.

A. Entropy-based evaluation

Entropy is used to measure both the pixel uniformity in a region and the complexity of overall segmentation using this evaluation. Given an image I with segmentation output S = R1, , Rn, a measure M is formalize. The

Supervised methods	Unsupervised methods
Need Ground Truth	Not need a Ground Truth
Ground Truth maybe ambiguous for a complex scenery	Avoid ambiguity in Ground Truth
No training phase	Explicitly allowed a training phase
Cannot find the optimal parameterization automatically	Can find the optimal parameterization automatically

Table VII
COMPARISON BETWEEN SUPERVISED AND UNSUPERVISED METHODS

objective of image segmentation normally is partitioning an image into regions that is homogeneous. Most of the algorithms balance the region homogeneity with number of regions and differences between adjacent regions. Two entropy measurements are taken during the evaluation:

- Region Entropy - a measure of region homogeneity.
- Layout Entropy - a measure of number of regions.

1) Region Entropy: The entropy of region i is given as below:

$$H(R_i) = -\sum_x \frac{N_i(x)}{|R_i|} \log \frac{N_i(x)}{|R_i|} \tag{17}$$

- $H(Ri) =$ Entropy of Region i
- $Ni(x) =$ No. of pixel in region i, Ri with value x

The expected region entropy for an image I can be obtained as a weighted sum of individual region entropies:

$$H_r(I) = \sum_i \frac{|R_i|}{|l|} H(R_i) \tag{18}$$

Lesser bits are required for encoding if a region has more feature points with same values and thus the entropy is lower. Hence, if an image contains many small regions, it is likely to have lower entropy. If each pixel is its own region, the expected entropy will be zero. The region entropy is biased towards over-segmentation. It can be balanced by layout entropy which is described below.

2) Layout Entropy: Layout entropy is a measure of the number of bits used to label the region to which each pixel belongs. When the number of the regions increases, the expected region entropy is decreased and the layout entropy is increased. Layout entropy is biased towards under-segmentation while the region entropy is biased towards over-segmentation. Its formula is as shown below:

$$H_l(I) = -\sum_i \frac{|R_i|}{|I|} \log \frac{|R_i|}{|I|} \tag{19}$$

Since layout entropy has the effect that is opposite to the region entropy, they can be combined to balance the evaluation measures. The resulting entropy measure is given as below:

$$E = H_r(I) + H_l(I) \tag{20}$$

VI. COMPARISON

Table VII presents a comparison between supervised evaluations and unsupervised evaluations.

VII. CONCLUSION

In this paper, we gave an overview of several methods permitting the evaluation of segmentation. We categorized these methods in two fields: supervised and unsupervised methods. Supervised methods are using Ground-Truth as absolute value while unsupervised methods are computed without any absolute knowledge. Supervised methods were composed of evaluation metrics based, local and global consistency error and Huang and Dom evaluation measure. Unsupervised methods were composed of entropy based evaluation and more precisely region entropy and layout entropy.

REFERENCES

[1] K. McGuinness and N. E. S. O'Connor, "Image segmentation, evaluation, and applications," Ph.D. dissertation, Dublin, Ireland, 2010. [Online]. Available: http://doras.dcu.ie/14998/

[2] M. Meilva, "Comparing clusterings: an axiomatic view," in *ICML '05: Proceedings of the 22nd international conference on Machine learning*. New York, NY, USA: ACM, 2005, pp. 577–584.