14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, September 4-8, 2006, copyright by EURASIP

SUPERVISED EVALUATION OF SYNTHETIC AND REAL CONTOUR SEGMENTATION RESULTS

S. Chabrier, H. Laurent, C. Rosenberger

Laboratoire Vision et Robotique – UPRES EA 2078
ENSI de Bourges – Université d'Orléans
10 boulevard Lahitolle, 18020 Bourges Cedex , France
email: helene.laurent@ensi-bourges.fr

Y.-J. Zhang

Department of Electronic Engineering
Tsinghua University
Beijing 100084, China
email: zhang-yj@tsinghua.edu.cn

ABSTRACT

This article presents a comparative study of 14 supervised evaluation criteria of image segmentation results. A preliminary study made on synthetic segmentation results allows us to globally characterise the behaviours of the selected criteria. This first analysis is then completed on a selection of 300 real images extracted from the ©Corel database. Ten contour segmentation methods based on threshold selection are used to generate real segmentation results and various situations corresponding to under- and over-segmentation. Experimental results permit to reveal the advantages and limitations of the studied criteria face to various situations.

1. INTRODUCTION

We present in this article an overview of the most common supervised evaluation criteria of image segmentation methods based on contour detection [1], [2], [3], [4], [5]. These criteria are based on the computation of a dissimilarity measure between a segmentation result and a ground truth as illustrated in Figure 1.

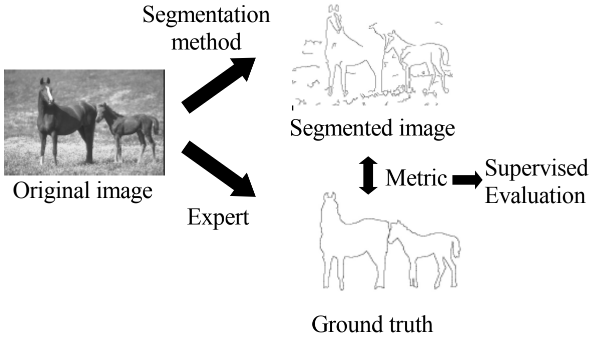

Figure 1 – Supervised evaluation procedure of a segmentation result

The ground truth can be determined by an expert. In this case, the obtained performance characterisation is inherently dependent on the confidence in the ground truth. Such a situation is widely encountered for medical applications when very specific knowledge is required. Whatever, it is often difficult to have a large amount of valued data at our disposal. In order to statistically evaluate the criteria performances, the ground truth can also be set during the generation of synthetic images. In this case, the ground truth is objective and known with an extreme precision but not necessarily representative.

After presenting the tested criteria, we first studied their behaviours on synthetic segmentation results corresponding to several degradations of a known ground truth. Then, we tested the chosen criteria on a selection of real images extracted from the ©Corel database for which manual segmentation results provided by experts are available. Contrary to synthetic cases, this basis allows to process the diversity of the possible encountered situations. It indeed contains images corresponding to different application fields such as aerial photography or landscape images [6]. The produced experimental results allow to compare the efficiency of the various supervised evaluation criteria.

2. STUDIED CRITERIA

Let I_{ref} be the reference contours corresponding to the ground truth, I_F the detected contours obtained through the segmentation result of image I.

2.1 Detection errors

Different index have been initially proposed to measure three detection errors [1]. The over-detection error (*ODE*) corresponds to detected contours which do not coincide with I_{ref}. The under-detection error (*UDE*) corresponds to pixels of I_{ref} which have not been detected. Lastly, the localisation error (*LE*) computes the distance between the misclassified pixels and the nearest pixels of I_{ref}. A good segmentation result should simultaneously minimise these three types of error.

2.2 L_q distances and divergence measures

Well known L_q distances ($q \geq 1$) can be used to compare two contour maps. We studied these distances for $1 \leq q \leq 4$ (L_1, L_2, L_3, L_4), the classical root mean squared error being obtained for q=2. Three distances, issued from probabilistic interpretation of images, complete the tests: the Küllback and Bhattacharyya (D_{Ku} and D_{Bh}) distances and the "Jensen-like" divergence measure (D_{Je}) based on Rényi entropies [2]. These measures provide a global comparison between two segmentation results but can express in a very inaccurate way some encountered deformations. They are for example unable to take into account possible geometrical shifts.

2.3 Hausdorff distance

The Hausdorff criterion (*HAU*) measures the distance between two pixel sets [3]. If $HAU(I_F,I_{ref})=d$, this means that all the pixels belonging to I_F are not further than d from some pixels of I_{ref} and *vice versa*. This measure is theoretically very interesting but seems to be noise sensitive.

2.4 Pratt figure of merit

In [4], Pratt proposes an empirical measure for two pixel sets comparison (*PRA*). Even if this measure is one of the most commonly used, it has no theoretical proof. The known drawbacks are that this criterion is not symmetrical, is sensitive to over-segmentation and localisation problems and does not express under-segmentation or shape errors.

2.5 Odet criteria

Different measurements have been proposed in [5] to estimate various errors in binary segmentation results. Amongst them, two divergence measures seem to be particularly interesting. The first one (OC_o) evaluates the divergence between the over-segmented pixels and the reference contour. The second one (OC_u) estimates the divergence between the under-segmented pixels and the computed contour. These criteria take into account the relative position for the over- and under-segmented pixels. A threshold d_{TH}, which has to be set according to each application precision requirement, permits to differently take into account the pixels with regard to their distance from the reference contour. These criteria also allow to differently weight the estimated contour pixels that are close to the reference contour and those having a distance to the reference contour close to d_{TH}.

3. PRELIMINARY STUDY

In order to study the behaviours of the previously presented criteria face to different perturbations, we first generated synthetic segmentation results corresponding to several degradations of the ground truth. The used ground truth is composed of five components: the central ring and the four external contours (see figure 2). The tested perturbations are the following:

- under-segmentation: one or several components of the ground truth are missing,
- over-segmentation affecting the complete image: noisy ground truth with impulsive noise (probability from 0.1 to 50%),
- over-segmentation affecting the contour area: dilatation of the contours with a width from 1 to 5 pixels,
- localisation error: synthetic segmentation results obtained by contour shifts from 1 to 5 pixels.

Different examples of the considered perturbations are presented in figure 2.

Figure 3 presents the evolution of three criteria (*ODE*, L_1, *HAU*) face to under-segmentation. 32 synthetic segmentation results corresponding to a more or less important under-segmentation have been created. *ODE* is equal to zero whatever case is considered. As *ODE* only measures over-segmentation, it equivalently grades a segmentation result missing one or several components. OC_o has the same behaviour. L_1 presents different stages allowing to gradually penalise under-segmentation. This behaviour corresponds to the expected one and the majority of the criteria evolves in that way (*UDE, LE*, L_2, L_3, L_4, D_{Ku}, D_{Bh}, D_{Je}, *PRA*). *HAU* also presents a graduate evolution but seems to suffer of a lack of precision. It equivalently grades some segmentation results even if the number of detected components is completely different.

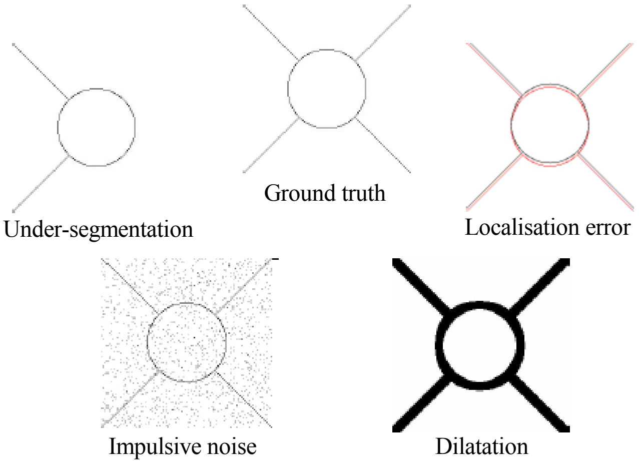

Figure 2 – The ground truth and some perturbations

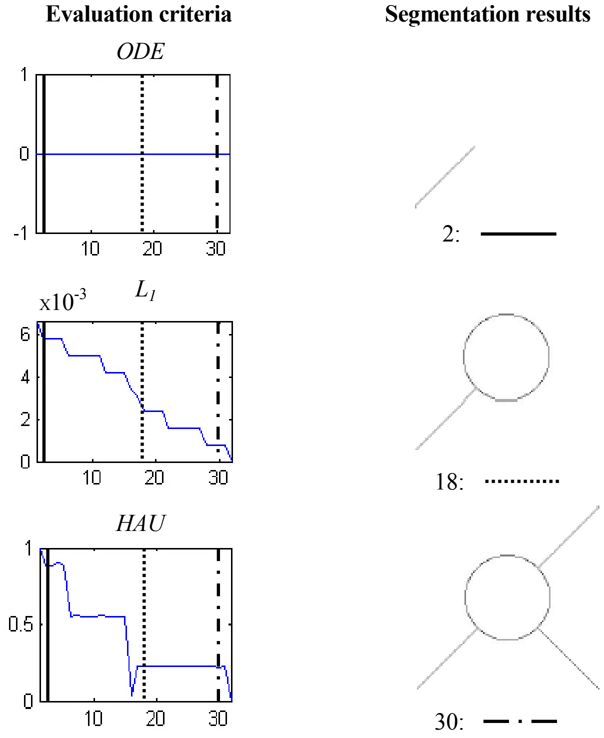

Figure 3 – Three evaluation criteria face to under-segmentation (x-axis : 32 synthetic segmentation results; y-axis : criterion value)

Figure 4 presents the evolution of three criteria (*UDE*, D_{Bh}, *PRA*) face to over-segmentation corresponding to the presence of impulsive noise. 13 synthetic segmentation results corresponding to a more or less important presence of noise have been created. As *UDE* only measures under-segmentation, it equivalently grades segmentation results with small or high presence of noise. OC_u have the same behaviour. D_{Bh} really penalises the over-segmentation only when it reaches a high level. *ODE, LE,* L_1, L_2, L_3, L_4, D_{Ku}, D_{Je} have the same kind of behaviour. *PRA* permits to penalise the presence of impulsive noise as soon as it appears. This criterion is the only one presenting this behaviour that is closer to the human decision: an expert will notice the presence of noise even for a small proportion and will immediately penalise it. On the other hand, he will not grade very differently too noisy segmentation results.

Concerning over-segmentation due to the dilatation of contours, except *UDE* and OC_u which are equal to zero whatever case is considered, the other criteria present the same behaviour. At last, all the criteria evolve in a similar way face to localisation error.

Finally, we can notice that, on our test set, the majority of the considered criteria have a similar behaviour, true to human judgement. The criteria which revealed themselves interesting are: *LE,* L_2, L_3, L_4, D_{Ku}, D_{Bh}, D_{Je}, *PRA*. It seems to emerge from this preliminary study that *PRA* gives the more discriminating decision.

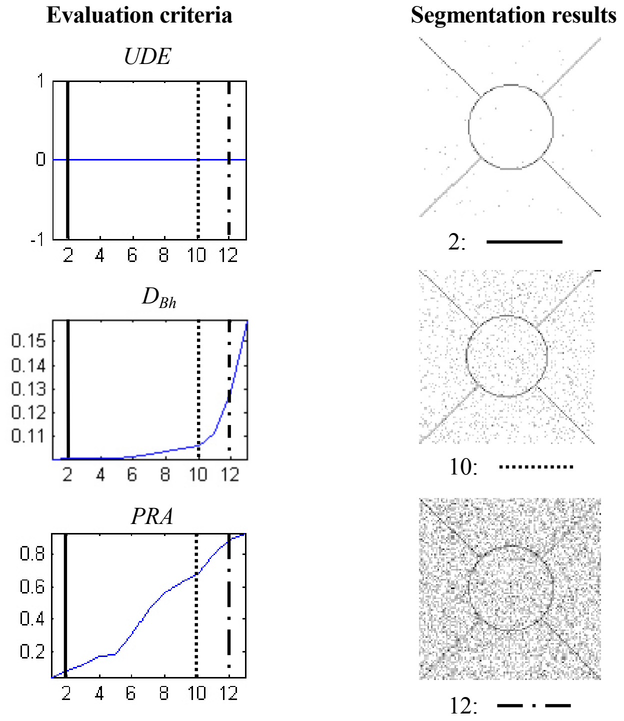

Figure 4 – Three evaluation criteria face to over-segmentation with impulsive noise (x-axis : 13 synthetic segmentation results; y-axis : criterion value)

4. TESTS ON REAL IMAGES

After this preliminary study, we studied the supervised evaluation criteria on real segmentation results obtained from natural images extracted from the ©Corel database [6].

4.1 Image database

The database used for our tests contains 300 real images extracted from the ©Corel database for which manual segmentations provided by experts are available. Figure 5 presents some of these images and different ground truths manually made by experts. We can notice that these reference segmentation results can be quite dissemblable. We then decided to make a fusion of the different expert ground truths in order to obtain a more relevant one.

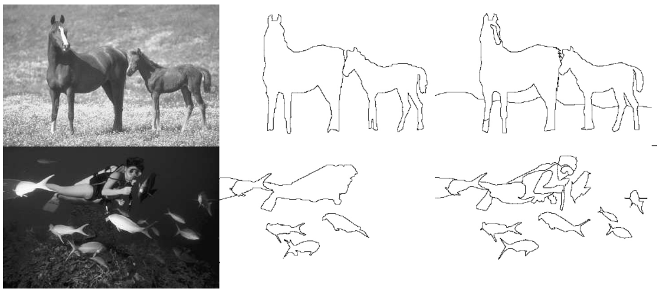

Figure 5 – Examples of real images extracted from the ©Corel database and corresponding experts ground truths

4.2 Segmentation results

We used 10 segmentation algorithms based on threshold selection to generate real segmentation results [7]:

- first moment matrix,
- second moment matrix,
- color/texture gradients,
- brightness/texture gradients,
- gradient multi-scale magnitude,
- color gradient,
- texture gradient,
- brightness gradient,
- gradient magnitude,
- Canny filter.

These filters generate fuzzy contour maps as shown in figure 6.

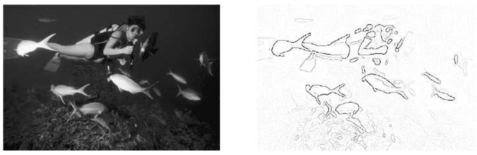

Figure 6 – Example of the fuzzy contour map obtained for one original image of the ©Corel database with the Canny filter

As we need binary contour maps, we threshold the obtained segmentation results. To choose the threshold value (*TH)*, we empirically established the following normalized formula, searching for a good compromise between under- and over-segmentation and taking into account the common pixels in the fused ground truth and the segmentation result:

$$TH(I,FGT,Seg)=\max_{\alpha\in 1..256}\left(\frac{Card(Seg(\alpha)\cap FGT)}{Card(FGT)}-\frac{Card(Seg(\alpha))}{2NBP_I}+\frac{1}{2}\right)$$

where *I* corresponds to the original image, *Seg* to the segmentation result to be binarised, *FGT* to the fused ground truth and where *Card(X)* is the number of contour pixels of *X* and NBP_I the number of pixels of *I*. Given this threshold characterising the compromise, we afterwards defined two situations simulating under- and over-segmentation. In the first case, 30% of pixels are missing. In the second one, 30%

of extra pixels are added. Figure 7 presents two original images, their fused ground truths and three segmentation results (under-segmented, compromise and over-segmented) obtained with the Canny filter.

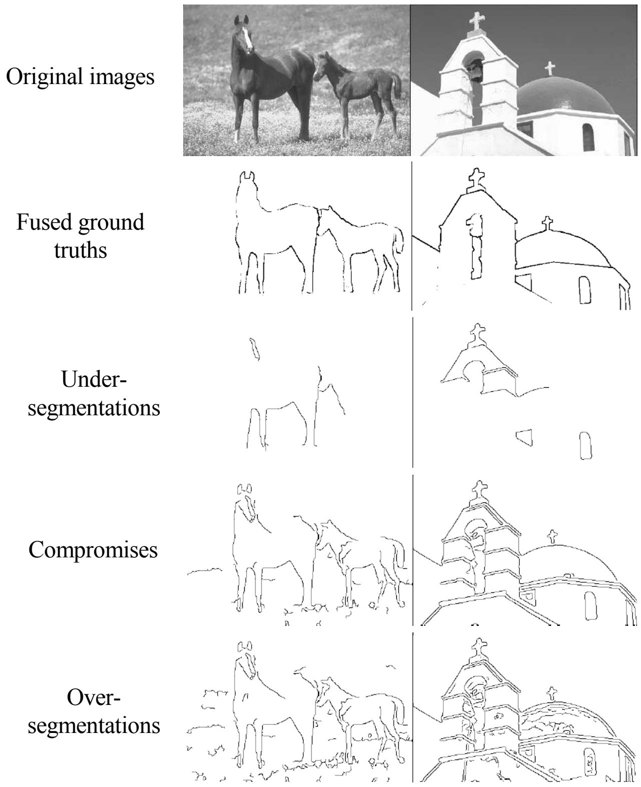

Figure 7 – Examples of original images, corresponding fused ground truths and segmentation results (under-segmented, compromise and over-segmented) obtained with the Canny filter

4.3 Experimental results

Table 1 presents the average sorting of the three segmentation results (under-segmented, compromise and over-segmented). These statistics highlight, with rank 1, the preferred situation for each criterion.

As a result of their definitions, ODE / OC_o and UDE / OC_u prefer the under- and the over-segmentation respectively. We can notice that for the criteria LE, L_2, L_3, L_4, D_{Ku}, D_{Bh}, D_{Je}, under-segmentation is elected by a majority. The only one criterion which allows to detect the compromise as being the best result is PRA.

5. CONCLUSION

In this article, we presented a comparative study of 14 supervised evaluation criteria of segmentation results. These criteria are based on the computation of a dissimilarity measure between a segmentation result and a ground truth. Even if, for real applications, this ground truth is not easily accessible and depends in an important way from the expert sensitivity, lots of medical applications rely on the assumption that a ground truth exists. The majority of the selected criteria, even if they have correct behaviours according to human judgement, penalise in a more important way over-segmentation face to under-segmentation. One criterion stands out from this study: Pratt figure of merit (*PRA*). With real and synthetic images corresponding to different application fields, this criterion revealed itself as the most effective. Concerning real images it allows to choose in 89% cases the compromise as being the best segmentation result.

Criteria	Under-segmentation	Compromise	Over-segmentation
ODE	**1.1**	2.0	3.0
UDE	2.8	1.9	**1.0**
LE	**1.1**	2.0	3.0
L_1	**1.1**	2.0	3.0
L_2	**1.1**	2.0	3.0
L_3	**1.1**	2.0	3.0
L_4	**1.1**	2.0	3.0
D_{Ku}	**1.1**	2.0	3.0
D_{Bh}	**1.1**	2.0	3.0
D_{Je}	**1.1**	2.0	3.0
HAU	1.4	2.0	2.9
PRA	1.7	**1.3**	2.8
OC_o	**1.3**	1.9	2.7
OC_u	2.8	1.9	**1.0**

Table 1 – Mean ranking of each segmentation result for the different criteria (rank 1 corresponds to the optimal one)

REFERENCES

[1] S. Chabrier, "Contribution à l'évaluation de performances en segmentation d'images," Thèse de doctorat, Université d'Orléans, décembre 2005.

[2] O. Michel, R.G. Baraniuk, and P. Flandrin, "Time-frequency based distance and divergence measures," in *Proc. TFTS'94*, 1994, pp. 64-67.

[3] M. Beauchemin, KP.B. Thomson, and G. Edwards, "On the Hausdorff distance used for the evaluation of segmentation results," *CJRS*, 24(1), pp. 3-8, 1998.

[4] W. Pratt, O. D. Faugeras and A. Gagalowicz, "Visual discrimination of stochastic texture fields," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 11(8), pp 796-804, 1978.

[5] C. Odet, B. Belaroussi and H. Benoit-Cattin, "Scalable Discrepancy Measures for Segmentation Evaluation," in *Proc. ICIP*, 2002, pp. 785-788.

[6] D. Martin, C. Fowlkes, D. Tal and J. Malik, "A Database of Human Segmented Natural Images and its Application to Evaluating Segmentation Algorithms and Measuring Ecological Statistics," *8th International Conference Computer Vision*, 2001, pp. 416-423.

[7] D. Martin, C. Fowlkes and J. Malik, "Learning to detect natural image boundaries using local brightness, color and texture cues," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26(1), 2004.