Evaluation Protocol for Localization Metrics
Application to a Comparative Study

Baptiste Hemery[1,⋆], Hélène Laurent[1], Christophe Rosenberger[2], and Bruno Emile[1]

[1] Institut Prisme
ENSI de Bourges - Université d'Orléans
88 boulevard Lahitolle, 18020 Bourges - France
baptiste.hemery@ensi-bourges.fr
[2] Laboratoire Greyc
ENSICAEN - Université de Caen - CNRS
6 boulevard du Maréchal Juin, 14000 Caen - France

Abstract. Localization metrics permit to quantify the correctness of object detection in an image interpretation result. This paper deals with the definition of a protocol in order to evaluate the behavior of localization metrics. We first define some properties that metrics should verify and create a synthetic database that enables to verify those properties on different metrics. After presenting the tested localization metrics, the results obtained following the proposed protocol are exposed. Finally, some conclusions and perspectives are given.

1 Introduction

Image processing deals with lots of methods from image acquisition (with camera, webcam, satellite...) to image interpretation. Image interpretation consists in extracting information about objects present in an image. Among these information, the automatic localization of an object is a great challenge. As this information is important for many applications, it is required for this localization to be as precise as possible.

In order to evaluate localization algorithms, several research competitions were created such as the Pascal VOC Challenge [1] or the French Robin Project [2]. Given a ground truth, these competitions need a reliable metric to evaluate and compare results obtained by different localization algorithms.

The object localization in an image can be done in many ways: center of the object, bounding box, contour or pixels binary mask. If the localization using the center of the objet can be easily evaluated by the Euclidean distance for example, it is not so easy for the other types of localization. By the way, many metrics have been created to evaluate a localization result obtained via a bounding box, a contour or a pixel binary mask. As examples, the Pascal VOC challenge[1] [1] uses a metric based on pixel binary mask, whereas the Robin Project[2] [2]

⋆ The author would like to thank the French Research Ministry for their financial help.

[1] http://www.pascal-network.org/challenges/VOC/

[2] http://robin.inrialpes.fr/

A. Elmoataz et al. (Eds.): ICISP 2008, LNCS 5099, pp. 273–280, 2008.

created three metrics based on a bounding box. Martin [3] created two metrics based on masks in order to assess segmentation results manually defined done by different persons. Odet [4] also created two metrics, based on contour, in order to evaluate segmentation result. Many other proposals can be found in the literature [5,6,7,8].

This paper aims to define a protocol to evaluate the reliability of a localization metric. This paper is divided in two parts: the first one deals with the definition of required properties for the metrics and the creation of a synthetic database used for the evaluation. The second part presents some tested metrics and give the obtained results. Finally, we conclude and give some perspectives of this study.

2 Evaluation Protocol

A way to evaluate a localization metric is to check if this metric verifies some specific properties. To achieve this goal, we suggest to use the principle of the supervised evaluation of localization, based on a ground truth and a localization result. The metric provides a score corresponding to the coherence between these two images. As we want to verify the properties of metrics, we propose to work in a totally controlled environment using two synthetic images. The first one corresponds to the ground truth and the second one, corresponding to a simulated localization result, is obtained by altering the ground truth (see Fig.1). As we control the alteration of the ground truth, we can study the evolution of the score given by the localization metric and verify if the metric has the expected properties.

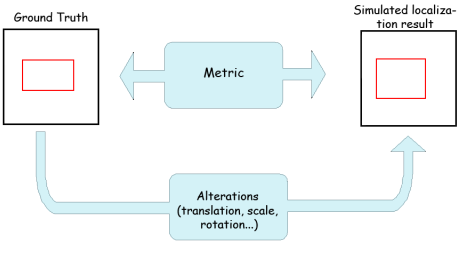

Fig. 1. Protocol Principle

2.1 Properties

We want to verify if a chosen localization metric has good performances regarding to the following properties:

1. Strict Monotony: a metric should penalize the results the more they are altered,
2. Symmetry: a metric should equally penalize two results with the same alteration, but in opposite directions,

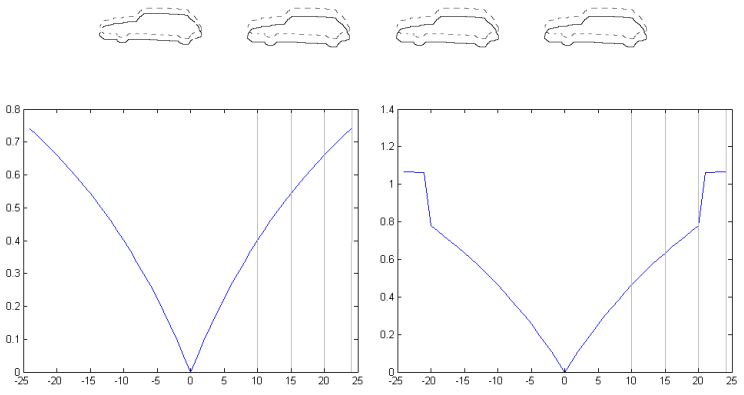

Fig. 2. Examples of localization metrics evaluation: both metrics are symmetric, the first metric is also strictly monotonous and uniformly continuous

3. Uniform Continuity: a metric should not have an important gap between two close results,
4. Topological dependence: a metric result should depend on the size or the shape of the localized object.

We can see on Fig.2 four cases of alteration of the same ground truth representing a car. This alteration consists in a translation along the vertical axis of 10, 15, 20 and 24 pixels. The two figures below show the evaluation results of similarly altered ground truths by two different metrics. The curves represent the value of the metric in function of the translation along the vertical axis. We can see that both metrics penalize similarly a translation of x pixels and $-x$ pixels. The first metric is monotone and continuous, the result is increasing with the translation and there is no gap. We can see on the second metric that there is a gap when the car is translated of 20 pixels, so the metric is not continuous. Moreover, we can see that after the gap, the metric equally penalizes all alterations, so the metric is not strictly monotonous.

These tested properties have been intuitively chosen and we plan to do a subjective study in order to confirm the importance of these properties for a localization metric evaluation considering a human expect.

2.2 Creation of a Synthetic Database

In order to verify the previously mentioned properties, we need a large amount of images couples corresponding to the ground truth and the simulated localization result. In order to create this database, we considered 16 ground truths that can be seen in Fig.3. We used 8 ground truths representing a bounding box with different sizes and shapes. We also consider the case where a ground truth is near the border of the image. We also created 8 other ground truths

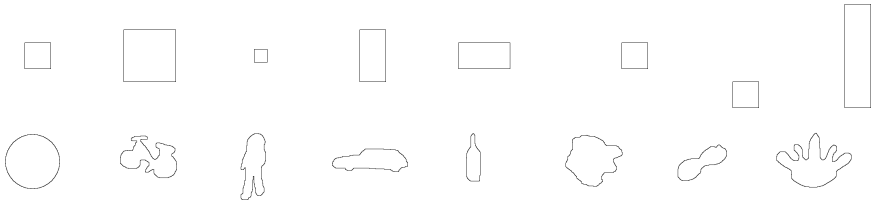

Fig. 3. Ground truths used for the creation of the database

corresponding to real objects that can be used in image interpretation: a bike, a car, a man... Those images are composed of 256 by 256 pixels.

In order to verify if the metrics have the required properties listed before, we used different alterations to create synthetic localization results. We used 4 alterations: translation, scale, rotation and perspective. The translation depends on two parameters: x and y. The parameter x describes a translation along the vertical axis and the y parameter a translation along the horizontal axis. Both parameters evolve between -24 pixels and +24 pixels which leads to 2.400 simulated localization results for one ground truth. The scale alteration depends on two parameters too. The parameter x denotes a scale along the vertical axis and the y parameter a scale along the horizontal axis. Those parameters evolve also between -24 pixels and +24 pixels. A negative value corresponds to a downscaling, whereas a positive value denotes an upscaling. We obtain 2.400 simulated localization results per ground truth. The rotation depends on only one parameter d, corresponding to the angle of rotation in degrees. The parameter d evolves between -90 °and +90 °. This leads to 180 simulated localization results for each ground truth.

The perspective alteration depends on two parameters. The parameter x corresponds to a perspective alteration along the vertical axis and the y parameter a perspective alteration along the horizontal axis. Those parameters evolve also between -24 pixels and +24 pixels. We obtain 2.400 simulated localization results per ground truth.

We can see on Fig.4 examples of altered ground truth. We finally obtain a total of **118.080** synthetic localization images.

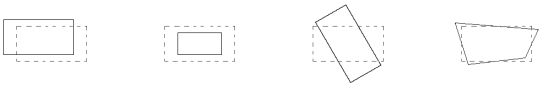

Fig. 4. Examples of altered ground truth

3 Comparative Study

We used this protocol to realize a comparative study of existing localization metrics. Several well know metrics, extracted from the state of the art, were computed on the database allowing to obtain a score for each metric depending of the previously mentioned properties.

3.1 Metrics

There are different types of outputs for a localization algorithm. We present, in the following paragraphs, examples of each type.

The most common one consists in the object localization by a couple of points representing a bounding box. The Robin project [2] aims at evaluating localization and recognition algorithms providing this type of information. In order to achieve this goal, three metrics have been developed to evaluate localization result described by a bounding box:

$$RobLoc(Z_l, Z_{gt}) = \frac{2}{\pi} arctan(max(\frac{|x_l - x_{gt}|}{w_{gt}}, \frac{|y_l - y_{gt}|}{h_{gt}})) \quad (1)$$

$$RobCom(Z_l, Z_{gt}) = \frac{|\mathcal{A}_l - \mathcal{A}_{gt}|}{max(\mathcal{A}_l, \mathcal{A}_{gt})} \quad (2)$$

$$RobCor(Z_l, Z_{gt}) = \frac{2}{\pi} arctan(|\frac{h_l}{w_l} - \frac{h_{gt}}{w_{gt}}|) \quad (3)$$

where Z_l is the output of the localization algorithm, $\{x_l, y_l\}$ are the coordinates of the center of the bounding box, $\mathcal{A}_l$ is the area covered by the bounding box and $\{h_l, w_l\}$ are the height and width of the bounding box. The variables using $_{gt}$ correspond to the same measures for the ground truth. These three metrics evaluate different characteristics of the localization result: RobLoc evaluates the localization of the center of the bounding box, RobCor quantifies the ratio height/width of the bounding box and RobCom evaluates the size of the bounding box.

There are two other types of localization based on contour or pixels binary mask for representing the localized object. The mask, or region, aspect is, for example, used in the Pascal VOC Challenge [1]. A simple metric is then used to evaluate the localization of an object:

$$PAS(I_{gt}, I_l) = \frac{Card(I_{gt}^{Re} \cap I_l^{Re})}{Card(I_{gt}^{Re} \cup I_l^{Re})} \quad (4)$$

with I_l^{Re} corresponds to region pixels of the localized object, $I_{gt}^{Re} \cap I_l^{Re}$ corresponds to object pixels correctly localized and $I_{gt}^{Re} \cup I_l^{Re}$ corresponds to object pixels from the ground truth or from the localized object.

The contour aspect is often used for segmentation evaluation. For example, the Figure of Merit (FOM) proposed by Pratt [5] is an empirical distance between the image with the contour of the localized object I_l and the corresponding ground truth I_{gt}:

$$FOM(I_{gt}, I_l) = \frac{1}{MP} \sum_{k \in I_l^{Cont}} \frac{1}{1 + \alpha * d(k, I_{gt}^{Cont})^2} \quad (5)$$

where I_l^{Cont} are contour pixels of the localized object, MP corresponds to $Max\big(Card(I_{gt}^{Cont}), Card(I_l^{Cont})\big)$, α is a constant fixed at $\frac{1}{9}$ and $d(x, I) = min_{y \in I} d(x, y)$.

Table 1. List of metrics used in the comparative study

Metric	type	reference	Metric	type	reference
RobLoc	Box	[2]	DBad	Contour	[7,9]
RobCom	Box	[2]	ODI	Contour	[4,8]
RobCor	Box	[2]	UDI	Contour	[4,8]
ErrLoc	Contour	[7,8]	PAS	Mask	[1]
ErrSous	Contour	[7,8]	Hen1	Mask	[10]
ErrSur	Contour	[7,8]	Hen2	Mask	[10]
SNR	Contour	[9,11]	Yas1	Mask	[6,8]
RMS	Contour	[9,11]	Yas2	Mask	[6,8]
Lq	Contour	[9,11]	Yas3	Mask	[6,8]
DKu	Contour	[8,12]	Mar1	Mask	[3,8]
DBh	Contour	[8,12]	Mar2	Mask	[3,8]
DJe	Contour	[8,12]	Ham	Mask	[13,8]
DMoy	Contour	[8,14]	Haf1	Mask	[15]
DMoC	Contour	[8,14]	Haf2	Mask	[16]
FOM	Contour	[5,14]	Vin	Mask	[17,18]
DHau	Contour	[7,19]			

For this comparative study, we used 31 localization metrics listed in Tab.1. Part of those metrics were not created with the specific purpose of localization evaluation, but for the segmentation evaluation or image quality evaluation.

3.2 Experimental Results

The obtained results are presented in Tab.2. Some metrics depend on an additional parameter, like the distance Lq, DBad, ODI and UDI metrics. We tested these metrics with several values. For each metric, a score is attributed depending upon its properties for each alteration. For example, we attribute a star to one metric if its result to translation is strictly monotonic, a second star if it is symmetric... Therefore, the translation, rotation and perspective alterations maximum score is 5, the scale alteration maximum is 4 because we do not verify if results are symmetric. We also give a final score to each metric which is the sum of scores for each alteration. By the way, the maximum final score is 19.

We can see that metrics used for the Robin project do not have good performance. This comes from the fact that those metrics must be used together in order to correctly evaluate a localization result.

Some metrics based on a contour result: ErrLoc, ErrSous, ErrSur, SNR, RMS, and distances Lq, DBh, DKu, and DJe obtain low scores. This enables to conclude that those metrics should not be used for the evaluation of a localization result. The other metrics based on contour: DMoy, DMoC, FOM, DBad, ODI and UDI metrics obtain quite good results.

Last, all metrics based on pixels binary masks obtain good results. We can note that only the first metric of Martin obtains the maximal score of 19.

Table 2. Synthesis of obtained results

Metric	Type	Translation	Scale	Rotation	Perspective	Final score
RobLoc	Box	*****	—	—	–	5
RobCom	Box	—	***	***	**	8
RobCor	Box	—	***	***	****	10
ErrLoc	Contour	***	**	***	****	12
ErrSous	Contour	***	**	***	****	12
ErrSur	Contour	***	**	***	***	11
SNR	Contour	***	**	**	***	10
RMS	Contour	***	**	****	***	12
Lq,1	Contour	***	**	****	****	13
Lq,3	Contour	***	**	***	***	11
DKu	Contour	***	**	****	****	13
DBh	Contour	***	***	***	**	11
DJe	Contour	***	**	****	****	13
DMoy	Contour	*****	****	****	***	16
DMoC	Contour	*****	****	****	***	16
FOM	Contour	*****	****	*****	****	**18**
DHau	Contour	***	**	***	***	11
DBad,1	Contour	****	***	*****	***	15
DBad,2	Contour	****	***	****	***	14
DBad,3	Contour	****	**	*****	***	14
ODI,1	Contour	****	**	*****	***	14
ODI,2	Contour	****	**	*****	***	14
UDI,1	Contour	****	**	*****	****	15
UDI,2	Contour	****	**	*****	***	14
PAS	Mask	*****	****	****	*****	**18**
Hen1	Mask	—	****	—	—	4
Hen2	Mask	*****	****	****	*****	18
Yas1	Mask	*****	***	****	*****	17
Yas2	Mask	*****	***	****	*****	17
Yas3	Mask	*****	***	****	*****	17
Mar1	Mask	*****	****	*****	*****	**19**
Mar2	Mask	*****	***	*****	*****	**18**
Ham	Mask	****	***	****	*****	16
Haf1	Mask	****	**	****	*****	15
Haf2	Mask	****	***	****	*****	16
Vin	Mask	*****	****	****	*****	**18**

4 Conclusions

We proposed in this article a protocol that enables to study and compare localization metrics. This protocol allows the comparison of metrics using different representations of a localization result. The comparative study clearly shows that some metrics should not be used for the localization evaluation. Metrics based on pixels binary masks generally give better results than other metrics.

We now plan to realize a subjective evaluation of localization results. This will show the expected properties by a human for localization metrics and will enable to improve our protocol for the evaluation of a localization metrics.

References

1. Everingham, M., Zisserman, A., Williams, C., Van Gool, L., Allan, M., Bishop, C., Chapelle, O., Dalal, N., Deselaers, T., Dorko, G., et al.: The 2005 pascal visual object classes challenge (2005)
2. D'Angelo, E., Herbin, S., Ratiéville, M.: Robin challenge evaluation principles and metrics (2006)
3. Martin, D., Fowlkes, C., Tal, D., Malik, J.: A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: 8th Int'l Conf. Computer Vision, pp. 416–423 (2001)
4. Odet, C., Belaroussi, B., Benoit-Cattin, H. (eds.): Scalable Discrepancy Measures for Segmentation Evaluation, vol. 1 (2002)
5. Pratt, W., Faugeras, O.D., Gagalowicz, A.: Visual discrimination of stochastic texture fields. IEEE Transactions on Systems, Man, and Cybernetics 8(11), 796–804 (1978)
6. Yasnoff, W.A., Mui, J.K., Bacus, J.W.: Error measures for scene segmentation. Pattern Recognition 9, 217–231 (1977)
7. Baddeley, A.J.: An error metric for binary images. Robust Computer Vision, 59–78 (1992)
8. Chabrier, S.: Contribution à l'évaluation de performances en segmentation d'images. PhD thesis, Université d'Orléans (2005)
9. Wilson, D.L., Baddeley, A.J., Owens, R.A.: A new metric for grey-scale image comparison. International Journal of Computer Vision 24(1), 5–17 (1997)
10. Henricsson, O., Baltsavias, E.: 3-d building reconstruction with aruba: A qualitative and quantitative evaluation. In: Automatic Extraction of Man-Made Objects from Aerial and Space Images (II), pp. 65–76 (1997)
11. Coquin, D., Bolon, P., Chehadeh, Y.: Evaluation quantitative d'images filtrées. In: GRETSI 1997, vol. 2, pp. 1351–1354 (1997)
12. Basseville, M.: Distance measures for signal processing and pattern recognition. Signal Processing 18(4), 349–369 (1989)
13. Huang, Q., Dom, B.: Quantitative methods of evaluating image segmentation. In: International Conference on Image Processing (ICIP 1995), Washington DC, USA, vol. 3, pp. 53–56 (1995)
14. Peli, T., Malah, D.: A study of edge detection algorithms. Comp. Graphics Image (1982)
15. Hafiane, A.: Caractérisation de textures et segmentation pour la recherche d'images par le contenu. PhD thesis, Université de Paris-Sud XI (2005)
16. Hafiane, A., Chabrier, S., Rosenberger, C., Laurent, H.: A new supervised evaluation criterion for region based segmentation methods. In: Blanc-Talon, J., Philips, W., Popescu, D., Scheunders, P. (eds.) ACIVS 2007. LNCS, vol. 4678, pp. 439–448. Springer, Heidelberg (2007)
17. Vinet, L.: Segmentation et mise en correspondance de régions de paires d'images stéréoscopiques. PhD thesis, Université de Paris IX Dauphine (1991)
18. Cocquerez, J.P., Philipp, S.: Analyse d'Images: filtrage et segmentation. Masson (1995)
19. Beauchemin, M., Thomson, K.B., Edwards, G.: On the hausdorff distance used for the evaluation of segmentation results. Canadian Journal of Remote Sensing 24(1), 3–8 (1998)