Comparative study of localization metrics for the evaluation of image interpretation systems

Baptiste Hemery, Hélène Laurent, Bruno Emile, Christophe Rosenberger

Comparative study of localization metrics for the evaluation of image interpretation systems

Baptiste Hemery
École Nationale Supérieure d'Ingéneurs de Caen
Université de Caen Basse-Normandie, CNRS
Laboratoire GREYC
6 Boulevard du Maréchal Juin
14000 Caen, France
E-mail: baptiste.hemery@greyc.ensicaen.fr

Helene Laurent
Bruno Emile
ENSI de Bourges, Université d'Orléans
Institut PRISME
88 Boulevard Lahitolle
18020 Bourges, France

Christophe Rosenberger
École Nationale Supérieure d'Ingéneurs de Caen
Université de Caen Basse-Normandie, CNRS
Laboratoire GREYC
6 Boulevard du Maréchal Juin
14000 Caen, France

Abstract. *Image interpretation is particularly important in many real applications (video monitoring, biometrics, etc.). Due to the proliferation of image interpretation systems in the literature, their evaluation still remains a crucial stake. Among all the tasks in this field, the quality of object localization is often evaluated through an evaluation metric. We propose to review these techniques and study their reliability. We first propose a generic definition of a localization algorithm. Then, different state of the art techniques to evaluate image interpretation results are detailed. Secondly, we focus on metrics that enable us to evaluate localization results. We propose a general methodology to analyze the behavior of an evaluation metric, considered here as a black box (its definition is not even supposed to be known). We define the properties that these metrics should fulfill. We then perform a comparative study of 33 localization metrics from the state of the art. Experimental results conducted on a large and significant image database permit us to determine metrics that should be used in the future for the evaluation of object localization results.* © 2010 SPIE and IS&T. [DOI: 10.1117/1.3446803]

1 Introduction

Image processing includes many steps from image acquisition (with camera, webcam, satellite, etc.) to image interpretation. Image interpretation consists in automatically extracting information of present objects in an image (detection of objects of interest, quantitative measure, etc.). As final goal of image processing, automatic image interpretation is a crucial step of the image processing chain. Among all the tasks in image interpretation, the automatic localization and recognition of an object in an image is still a great challenge.[1–3]

Whatever the foreseen application may be (biometric systems, medical imaging, video monitoring), the extracted information conditions the performances of the resulting process. It is required that this localization be as precise as possible and with correct recognition. Many algorithms have been proposed in the literature to achieve this task,[1–3] but it still remains difficult work to compare the performance of these algorithms, as we can see in Fig. 1.

To evaluate object detection algorithms, several research competitions have been created such as the Pascal VOC Challenge[4] or the French Robin Project.[5] These competitions are interesting in order to evaluate categorical object detection. Given a manually made ground truth, these com-

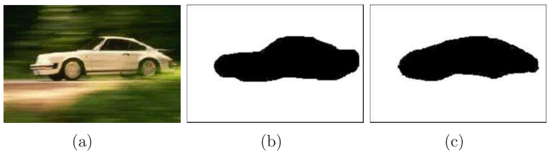


Fig. 1 Examples of localization results: (a) original image, (b) localization result 1, and (c) localization result 2.

Paper 09207RR received Oct. 26, 2009; revised manuscript received Apr. 23, 2010; accepted for publication Apr. 28, 2010; published online Jun. 16, 2010.

1017-9909/2010/19(2)/023017/21/$25.00 © 2010 SPIE and IS&T.

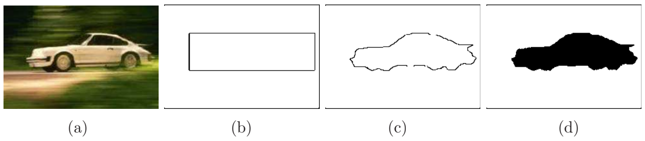


Fig. 2 Different localization representations: (a) original image, (b) bounding box, (c) contour, and (d) mask.

petitions use metrics to decide whether a region located by algorithms is correct or not. If the metrics used for these competitions appeal to everyone's common sense (good correspondence between the ratio height/width or the size of the detected bounding box and of the ground truth), none of them puts the same characteristic forward.

Moreover, many evaluation metrics initially proposed for various purposes, such as segmentation evaluation or image retrieval evaluation, can be found in the literature[6–9] and should reveal themselves relevant for localization evaluation. Simply the existence of all these metrics expresses the lack of localization algorithm evaluation. How can we choose a localization evaluation metric working on a single result to reliably evaluate an algorithm?

The goal of this work is to present a protocol that enables the comparison of localization evaluation metrics. The questions we want to answer in this work are the following. Which properties must fulfill an evaluation metric? Which metrics have the required properties and are the most suitable to quantify the quality provided by some localization algorithms? We present an evaluation protocol that enables us to study the reliability of localization metrics by defining the desired properties. We consider in this work an evaluation metric as a black box. Its mathematical definition is not supposed to be known. This protocol is then used for the comparative study of 33 localization metrics. Some illustrations of the efficiency of the studied metrics are given in Sec. 5. Finally, conclusions and perspectives of this study are given.

2 Evaluation of Localization Algorithms

The evaluation of image processing algorithms is not a recent problem.[6,10,11] To evaluate such algorithms, we can consider an image processing algorithm as a black box, with an original image and a set of parameters as inputs and a result as output (localization result in this case). The motivations of evaluation are multiple. First, it can be used to evaluate an algorithm to improve its performance during its development. Second, it allows a user to compare several algorithms to choose the best one for a specific application.

In the next section, we first give a general definition of the evaluation process in supervised and unsupervised contexts. We then focus on the definition of localization algorithms and their possible outputs. Finally, most of the localization metrics available in the literature are presented.

2.1 *Evaluation Methodologies*

The evaluation process can be either supervised or unsupervised. For an unsupervised evaluation, the evaluation process exploits the only input data used by the image interpretation algorithm, i.e., the original image, and gives a score of coherence that quantifies the possibility that the result given by the algorithm is correct. For a supervised evaluation, the evaluation process needs, in addition to the previous data, the ground truths corresponding to input images. By the way, this evaluation gives an adequacy score between results and the corresponding ground truths. This score is of course more reliable, but requires the definition of a ground truth for each input image.

Table 1 Notations.

BB_l	Coordinates of the bounding box in the localization result
A_l, h_l w_l, x_l, y_l	Area, height, width, and coordinates of the center of the localization result given by a bounding box, respectively
I_l	Localization result image
I	Common support of I_l and I_{gt}
I^c	Set of contour pixels of the image I
$I^c_{l\backslash gt}$	Set of contour pixels included in the image I_l but not in the image I_{gt}
I^r	Set of region/mask pixels of the image I
$I^{r(k)}$	Set of pixels from region/mask containing pixel k on the image I
$g_I(k)$	Gray level of pixel k in the image I
$d(k,I^c)$	Shortest distance from pixel k to set I^c, equal to $\min_{j\in I^c}[d(k,j)]$
Card$(\cdot)$	Cardinality of a set, number of elements in the set

The evaluation of an algorithm can be either the evaluation of a single result or a global evaluation. In the case where we want to evaluate the global performance of an algorithm, the objective is to study its general behavior considering different parameters, and to know what kind of alteration it is able to face to (illumination changes, presence of occlusion, etc.). The evaluation process permits us to study the influence of internal parameters of the algorithm and to optimize them. The goal is then to compute for each single image some performance measures on different results obtained by changing the parameters value. To realize a reliable evaluation, we need an evaluation database with enough images to be representative of most cases that the algorithm is used for.

To study the performances of localization algorithms, most of the organized competitions work in a supervised context and provide ground truth object annotations across all the proposed databases. We then focus on metrics that enable one to evaluate a single localization result in a supervised way.

2.2 *Localization Algorithms*

Localization consists in finding one or many objects of interest in an image and giving precisely their locations. This leads to the following definition of the localization:[5]

Table 2 List of used metrics.

Name	Metric	Representation	Formula	Parameters	Reference
Robin localization	$\mathrm{ROB}_{\mathrm{loc}}$	Box	$\mathrm{ROB}_{\mathrm{loc}}(BB_l, BB_{gt}) = \frac{2}{\pi} \arctan\left[\max\left(\frac{\lvert x_l - x_{gt}\rvert}{w_{gt}}, \frac{\lvert y_l - y_{gt}\rvert}{b_{gt}}\right)\right]$		5
Robin completeness	$\mathrm{ROB}_{\mathrm{com}}$	Box	$\mathrm{ROB}_{\mathrm{com}}(BB_l, BB_{gt}) = \frac{\lvert A_l - A_{gt}\rvert}{\max(A_l, A_{gt})}$		5
Robin correctness	$\mathrm{ROB}_{\mathrm{cor}}$	Box	$\mathrm{ROB}_{\mathrm{cor}}(BB_l, BB_{gt}) = \frac{2}{\pi} \arctan\left(\left\lvert \frac{h_1}{w_l} - \frac{h_{gt}}{w_{gt}} \right\rvert\right)$		5
	ErrLoc	Contour	$\mathrm{ErrLoc}(I_{gt}, I_l) = \frac{\mathrm{card}[(I^c_{gt\backslash l}) \cup (I^c_{l\backslash gt})]}{\mathrm{card}(I)}$		10 and 17
	ErrSous	Contour	$\mathrm{ErrSous}(I_{gt}, I_l) = \frac{\mathrm{card}(I^c_{gt\backslash l})}{\mathrm{card}(I^c_{gt})}$		10 and 17
	ErrSur	Contour	$\mathrm{ErrSur}(I_{gt}, I_l) = \frac{\mathrm{card}(I^c_{l\backslash gt})}{\mathrm{card}(I) - \mathrm{card}(I^c_{gt})}$		10 and 17
Signal noise ratio	SNR	Contour	$\mathrm{SNR}(I_{gt}, I_l) = \left[\frac{1}{\mathrm{card}(I)} \Sigma_{k \in I} \frac{gI_l(k)^2}{[gI_{gt}(k) - gI_l(k)]^2}\right]^{\frac{1}{2}}$		7 and 18
Root mean square	rms	Contour	$\mathrm{rms}(I_{gt}, I_l) = \left[\frac{1}{\mathrm{card}(I)} \Sigma_{k \in I} [gI_{gt}(k) - gI_l(k)]^2\right]^{\frac{1}{2}}$		7 and 18
Lq distance	L_q	Contour	$L_q(L_{gt}, I_l) = \left[\frac{1}{\mathrm{card}(I)} \Sigma_{k \in I} \left\lvert gI_l(k) - gI_{gt}(k) \right\rvert^g\right]^{\frac{1}{q}}$	$q \in \{1, 3\}$	7 and 18
Kullback distance	KUL	Contour	$\mathrm{KUL}(L_{gt}, I_l) = \frac{1}{\mathrm{card}(I)} \Sigma_{k \in I} [qI_l(k) - gI_{gt}(k)] * \log\left[\frac{gI_l(k)}{gI_{gt}(k)}\right]$		17 and 6
Bhattacharyya distance	BAH	Contour	$\mathrm{BHA}(I_{gt}, I_l) = -\log\left\{\frac{1}{\mathrm{card}(I)} \Sigma_{k \in I} [gI_{gt}(k) * gI_l(k)]^{\frac{1}{2}}\right\}$		17 and 6
Jensen distance	JEN	Contour	$\mathrm{JEN}(I_{gt}, I_l) = J\left(\frac{I_{gt} + I_l}{2}, I_{gt}\right); \; J(I_1, I_2) = H_\alpha(\sqrt{I_1 * I_2}) - \frac{H_\alpha(I_1) + H_\alpha(I_2)}{2}$ $H_\alpha(I_1) = \frac{1}{1 - \alpha} \log_2(\Sigma_{k \in I_1} [gI_1(k)]^\alpha)$	$\alpha = 3$	17 and 6
Mean distance	DMoy	Contour	$\mathrm{DMoy}(I_{gt}, I_l) = \frac{1}{\mathrm{card}(I^c_l)} \Sigma_{k \in I^c_l} d(k, I^c_{gt})$		19
Squared mean distance	DMoC	Contour	$\mathrm{DMoy}(I_{gt}, I_l) = \frac{1}{\mathrm{card}(I^c_l)} \Sigma_{k \in I^c_l} d(k, I^c_{gt})^2$		19
Figure of merit	FOM	Contour	$\mathrm{FOM}(I_{gt}, I_l) = \frac{1}{MP} \Sigma_{k \in I^c_l} \frac{1}{1 + \alpha * d(k, I^c_{gt})^2}$	$\alpha = \frac{1}{9}$	13 and 19
Hausdorff distance	HAU	Contour	$\mathrm{HAU}(I_{gt}, I_l) = \max[h(I_{gt}, I_l), h(I_l, I_{gt})]$ $h(I_1, I_2) = \max_{k_1 \in I^c_1} \min_{k_2 \in I^c_2} d(k_1, k_2)$		10 and 11
Baddeley distance	BAD	Contour	$\mathrm{BAD}(I_{gt}, I_l) = \left[\frac{1}{\mathrm{card}(I)} \Sigma_{k \in I^c_{gt} \cup I^c_l} \left\lvert d(k, I^c_{gt}) - d(k, I^c_l) \right\rvert^P\right]^{\frac{1}{P}}$ with $P \geqslant 1$	$P \in \{1, 2, 3\}$	10 and 7
Odet	ODI_n	Contour	$\mathrm{ODI}_n(I_{gt}, I_l) = \frac{1}{\mathrm{card}(I^c_{l\backslash gt})} \Sigma_{k \in I^c_{l\backslash gt}} \left[\frac{d(k, I^c_{gt})}{d_{Th}}\right]^n$	$n \in \{1, 2\}$; $d_{Th} = 5$	17 and 20
Odet	UDI_n	Contour	$\mathrm{UDI}_n(I_{gt}, I_l) = \frac{1}{\mathrm{card}(I^c_{gt\backslash l})} \Sigma_{k \in I^c_{gt\backslash l}} \left[\frac{d(k, I^c_l)}{d_{Th}}\right]^n$	$n \in \{1, 2\}$; $d_{Th} = 5$	17 and 20
Pascal	PAS	Mask	$\mathrm{PAS}(I_{gt}, I_l) = \frac{\mathrm{card}(I^r_{gt} \cap I^r_l)}{\mathrm{card}(I^r_{gt} \cup I^r_l)}$		4
Henricsson	HEN1	Mask	$\mathrm{HEN1}(I_{gt}, I_l) = \frac{\lvert \mathrm{card}(I^r_l) - \mathrm{card}(I^r_{gt}) \rvert}{\mathrm{card}(I^r_{gt})}$		21
Henricsson	HEN2	Mask	$\mathrm{HEN2}(I_{gt}, I_l) = \frac{\mathrm{card}(I^r_{l\backslash gt}) + \mathrm{card}(I^c_{gt\backslash l})}{\mathrm{card}(I^r_{gt})}$		21

Table 2 (*Continued.*)

Name	Metric	Representation	Formula	Parameters	Reference
Yasnoff	YAS1	Mask	$\mathrm{YAS1}(I_{gt},I_l,k)=100*\frac{\mathrm{card}[I_l^{r(k)}-\mathrm{card}(I_{gt\cap l}^{r(k)})]}{\mathrm{card}[I_l^{r(k)}]}$		17 and 22
Yasnoff	YAS2	Mask	$\mathrm{YAS2}(I_{gt},I_l,k)=100*\frac{\mathrm{card}[I_{gt}^{r(k)}-\mathrm{card}(I_{gt\cap l}^{r(k)})]}{\mathrm{card}(I)-\mathrm{card}[I_l^{r(k)}]}$		17 and 22
Yasnoff	YAS3	Mask	$\mathrm{YAS3}(I_{gt},I_l,k)=\frac{100}{\mathrm{card}(I)}\{\Sigma_{\alpha\in I_{l\backslash gt}^{r(k)}}d[\alpha,I_{gt}^{r(k)}]\}^{\frac{1}{2}}$		17 and 22
Martin, global consistency error	$\mathrm{MAR}_{\mathrm{gce}}$	Mask	$\mathrm{MAR}_{\mathrm{gce}}(I_{gt},I_l)=\frac{1}{\mathrm{card}(I)}\min[\Sigma_{k\in I}E(I_{gt},I_l,k),\Sigma_{k\in I}E(I_l,I_{gt},k)]$ $E(I_1,I_2,k)=\frac{\mathrm{card}[I_{1\backslash 2}^{r(k)}]}{\mathrm{card}[I_1^{r(k)}]}$		8 and 17
Martin local consistency error	$\mathrm{MAR}_{\mathrm{lce}}$	Mask	$\mathrm{MAR}_{\mathrm{ice}}(I_{gt},I_l)=\frac{1}{\mathrm{card}(I)}\Sigma_{k\in I}\min[E(I_{gt},I_l,k),E(I_l,I_{gt},k)]$		8 and 17
Hamming distance	HAM	Mask	$\mathrm{HAM}(I_{gt},I_l)=1-\frac{D_H(I_{gt},I_l)+D_H(I_l,I_{gt})}{2*\mathrm{card}(I)}$		17 and 23
Hafiane	HAF1	Mask	$\mathrm{HAF1}(I_{gt},I_l)=\eta\Sigma_{i,\mathrm{argmax}_j\ \mathrm{card}[I_{gt}^{r(i)}\cap I_l^{r(j)}]}\frac{\mathrm{card}[I_{gt}^{r(i)}\cap I_l^{r(j)}]}{\mathrm{card}[I_{gt}^{r(i)}\cup I_l^{r(j)}]}$ $\eta=\frac{N^*(I_{gt})}{N(I_l)}$ if $N(I_l)\geqslant N(I_{gt})$ or $\eta=\frac{1}{\log\left[\frac{N(I_{gt})}{N(I_l)}\right]}$ otherwise	$N(I)$ =number of objects in I	24
Hafiane	HAF2	Mask	$\mathrm{HAF2}(I_{gt},I_l)=\frac{M(I_{gt},I_l)+m\times\eta}{1+m}$ $m(I_{gt},I_l)=\Sigma_{j,\mathrm{argmax}_i\ \mathrm{card}[I_{gt}^{r(i)}\cap I_l^{r(j)}]}\frac{\mathrm{card}[I_{gt}^{r(i)}\cap I_l^{r(i)}]}{\mathrm{card}[I_{gt}^{r(i)}\cup I_l^{r(j)}]}\rho_j$	$m=0,2;$ $\rho_j=\frac{\mathrm{card}[I_l^{r(j)}]}{\mathrm{card}(I)}$	9
Vinet	VIN	Mask	$(I_{gt},I_l)=\mathrm{card}(I)-\Sigma_{C'}\mathrm{card}[I_l^{r(i)}\cap I_{gt}^{r(j)}]$		25 and 26
Pixel Precison	P_{px}	Mask	$P_{px}(I_{gt},I_l)=\frac{\mathrm{card}(I_{gt}^r\cap I_l^r)}{\mathrm{card}(I_l^r)}$		27
Pixel Recall	R_{px}	Mask	$R_{px}(I_{gt},I_l)=\frac{\mathrm{card}(I_{gt}^r\cap I_l^r)}{\mathrm{card}(I_{gt}^r)}$		27

$$\text{localization: } B,I\mapsto\{Z_i,\mu_i\}, \tag{1}$$

where B is a database with objects of interest if the algorithm is supervised and I is the original image. The localization result $\{Z_i,\mu_i\}$ is a list of potential localized objects, where Z_i is the localization of the object i and μ_i is the associated confidence in this result.

The localization result Z_i can have different representations: center of the object, bounding box, contour, or binary mask.[12] Considering for example the Pascal VOC Challenge, the ground truth information in each annotated image includes a bounding box for the object of interest and might also include a pixel segmentation mask or polygonal boundaries.[4]

The most common localization output consists of the object localization by a couple of points representing a bounding box. This type of localization is quite poor but is very easy to compute. A single point representing the center of the object can also be used, but this information is even poorer than the former and almost never used. The two other types of localization, based on contour or pixel binary mask for representing the localized object, are more precise but also make the creation of the annotated database more complicated. The contour localization consists of the use of 1 pixels that denote the frontier between the background and the object. The pixel binary mask, or region aspect, consists in using 1 pixels for the object of interest, whereas 0 pixels denote the background. We can see in Fig. 2 the three most commonly used representations of a localization result: a bounding box, a contour, and a binary mask.

2.3 *Metrics*

The supervised evaluation of a localization algorithm consists of comparing two images: the ground truth and the localization result. Each existing evaluation metric is dedicated to a type of representation of the localization result. In the following paragraphs, some examples of existing metrics are given for the different types of localization representations. The notations used in most cases for the localization result are gathered in Table 1. In the following paragraphs, variables using the subscript $\square_{gt}$ correspond to the same measures applied to the ground truth.

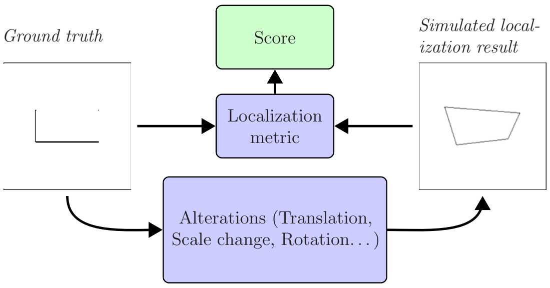

Ground truth
Score
Simulated localization result
Localization metric
Alterations (Translation, Scale change, Rotation. . .)

Fig. 3 Protocol principle.

2.3.1 *Bounding box metrics: Robin's metrics*

For the French Robin project,[5] which aims at evaluating localization and recognition algorithms providing bounding boxes as localization outputs, three metrics have been developed to evaluate a localization result:

$$\mathrm{ROB}_{\mathrm{loc}}(BB_l,BB_{gt}) = \frac{2}{\pi}\arctan\left[\max\left(\frac{|x_l-x_{gt}|}{w_{gt}},\frac{|y_l-y_{gt}|}{h_{gt}}\right)\right], \tag{2}$$

$$\mathrm{ROB}_{\mathrm{com}}(BB_l,BB_{gt}) = \frac{|\mathcal{A}_l-\mathcal{A}_{gt}|}{\max(\mathcal{A}_l,\mathcal{A}_{gt})}, \tag{3}$$

$$\mathrm{ROB}_{\mathrm{cor}}(BB_l,BB_{gt}) = \frac{2}{\pi}\arctan\left(\left|\frac{h_l}{w_l}-\frac{h_{gt}}{w_{gt}}\right|\right), \tag{4}$$

where BB_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. These three metrics evaluate different characteristics of the localization result: $\mathrm{ROB}_{\mathrm{loc}}$ evaluates the localization of the center of the bounding box, $\mathrm{ROB}_{\mathrm{com}}$ evaluates the size of the bounding box, and $\mathrm{ROB}_{\mathrm{cor}}$ quantifies the ratio height/width of the bounding box.

2.3.2 *Contour metric: Figure of merit metric*

Concerning the contour representation, several metrics have been proposed initially for segmentation evaluation. They can be easily extended to the localization evaluation. For example, the figure of merit (FOM) proposed by Pratt, Faugeras, and Gagalowicz[13] is an empirical distance between the image with the contour of the localized object I_l and the corresponding ground truth I_{gt}:

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

where

$$MP = \max[\mathrm{card}(I_{gt}^C),\mathrm{card}(I_l^C)], \tag{6}$$

and I_l^C are contour pixels of the localized object, α is a constant set to $1/9$ by the authors,[13] and $d(x,I) = \min_{y\in I} d(x,y)$.

2.3.3 *Region metric: Pascal and Martin's metrics*

The mask or region representation is, for example, used in the Pascal VOC Challenge.[4] A simple metric is then defined to evaluate the localization result of an object:

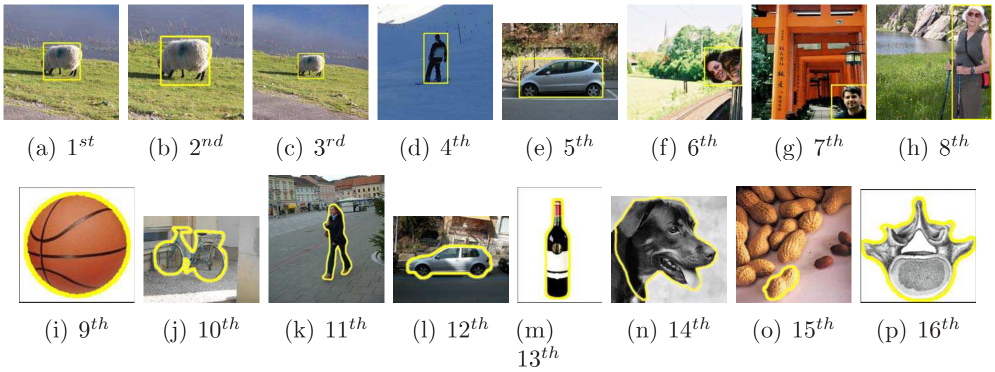
$$\mathrm{PAS}(I_{gt},I_l) = \frac{\mathrm{card}(I_{gt}^r \cap I_l^r)}{\mathrm{card}(I_{gt}^r \cup I_l^r)}, \tag{7}$$

where I_l^r corresponds to region pixels of the localized object, $I_{gt}^r \cap I_l^r$ corresponds to the object pixels correctly localized, and $I_{gt}^r \cup I_l^r$ corresponds to object pixels from the ground truth or from the localized object. This metric equals 1 when $I_{gt}^r \cap I_l^r = I_{gt}^r \cup I_l^r$, that is to say, when $I_{gt}^r = I_l^r$.

Two other metrics have been proposed by Martin *et al.*[8] for the evaluation of a localization result. These metrics work for several objects localized in a single image result. These metrics use the local refinement error between two images I_1 and I_2, defined as:

$$(I_1,I_2,k) = \frac{\mathrm{card}[I_{1\setminus 2}^{r(k)}]}{\mathrm{card}[I_1^{r(k)}]}, \tag{8}$$

where $r(k)$ corresponds to the region containing a pixel k. We can notice that this local error measure is not symmetric and only measures a refinement from image I_1 to image I_2. Martin *et al.* use this local refinement error to create two metrics called global consistency error (GCE) and local consistency error (LCE):


(a) 1st (b) 2nd (c) 3rd (d) 4th (e) 5th (f) 6th (g) 7th (h) 8th

(i) 9th (j) 10th (k) 11th (l) 12th (m) 13th (n) 14th (o) 15th (p) 16th

Fig. 4 Contour ground truths used for the creation of the database.

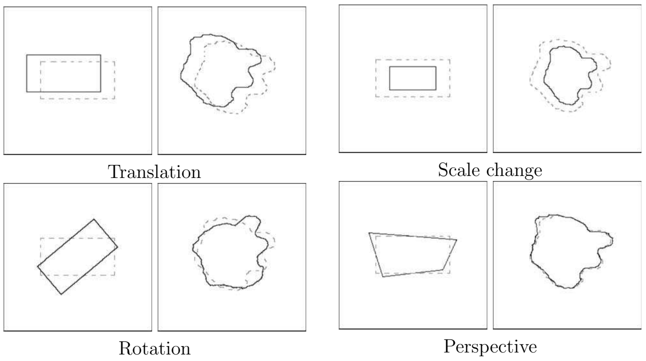


Fig. 5 Examples of alterations: dashed lines correspond to ground truths.

$$\mathrm{MAR}_{\mathrm{gce}}(I_{gt},I_l) = \frac{1}{\mathrm{card}(I)} \min\left[\sum_{k\in I}(I_{gt},I_l,k),\sum_{k\in I}(I_l,I_{gt},k)\right], \tag{9}$$

$$\mathrm{MAR}_{\mathrm{lce}}(I_{gt},I_l) = \frac{1}{\mathrm{card}(I)} \sum_{k\in I} \min[(I_{gt},I_l,k),(I_l,I_{gt},k)], \tag{10}$$

with I being the common support image of I_{gt} and I_l. We can notice that both metrics are symmetric. Moreover, it is clear that $\mathrm{MAR}_{\mathrm{gce}}$ is tougher than $\mathrm{MAR}_{\mathrm{lce}}$, since $\mathrm{MAR}_{\mathrm{gce}}$ forces all local refinement to be in the same direction (either from I_l to I_{gt} or from I_{gt} to I_l), whereas $\mathrm{MAR}_{\mathrm{lce}}$ allows refinement in both directions.

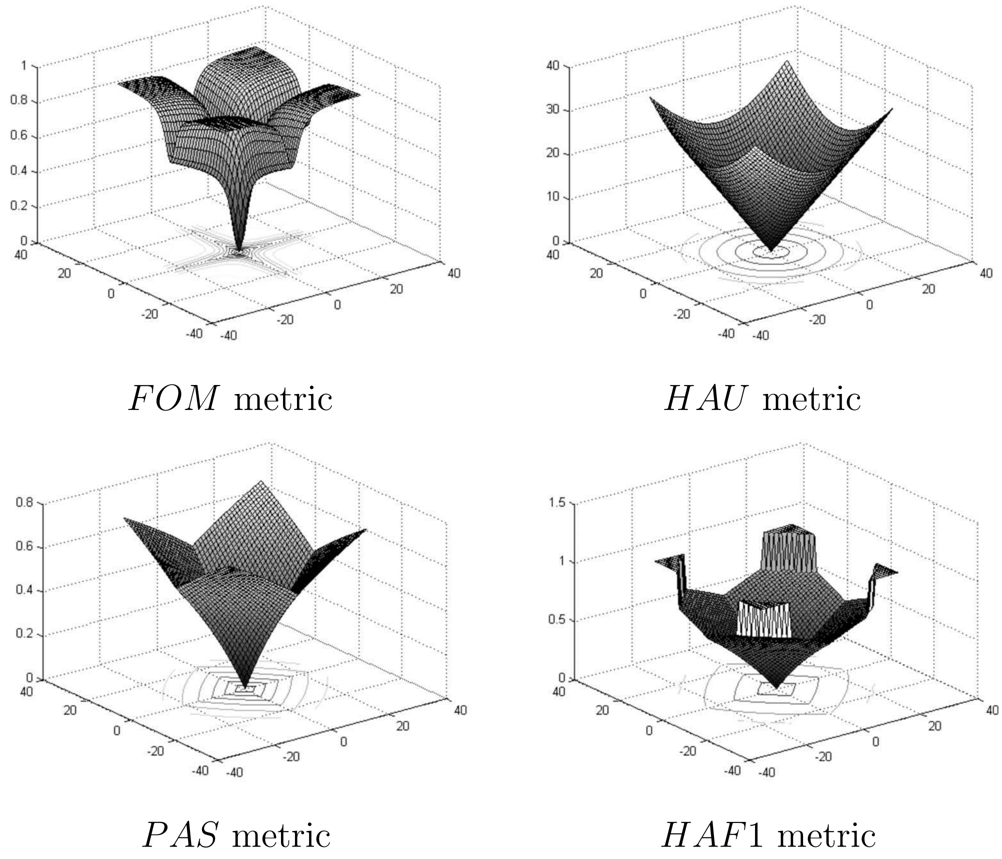
2.3.4 *Discussion*

Studied metrics are gathered in Table 2. As previously mentioned, some of those metrics have not been created with the specific purpose of localization evaluation, but for segmentation or image quality evaluation. As these metrics enable one to compare two images, we also included them in the comparative study.

We can also think about graph-based metrics (edit distance[14] or shock graph matching[15]) or moments features. Such metrics are used to compare detected and recognized shapes. By the way, they enable a comparison between highly different shapes and are robust to alterations. These metrics are then mostly used for the recognition step of an interpretation algorithm. They therefore do not correspond to the purpose of this study, which consists in quantifying the dissimilarity between a ground truth and a localization result.

3 Experimental Protocol

A way to quantify the reliability of a localization metric is to check if this metric verifies some specific properties. As all the considered metrics provide a score corresponding to the adequacy between a ground truth and a localization result, we propose to work in a totally controlled environ-



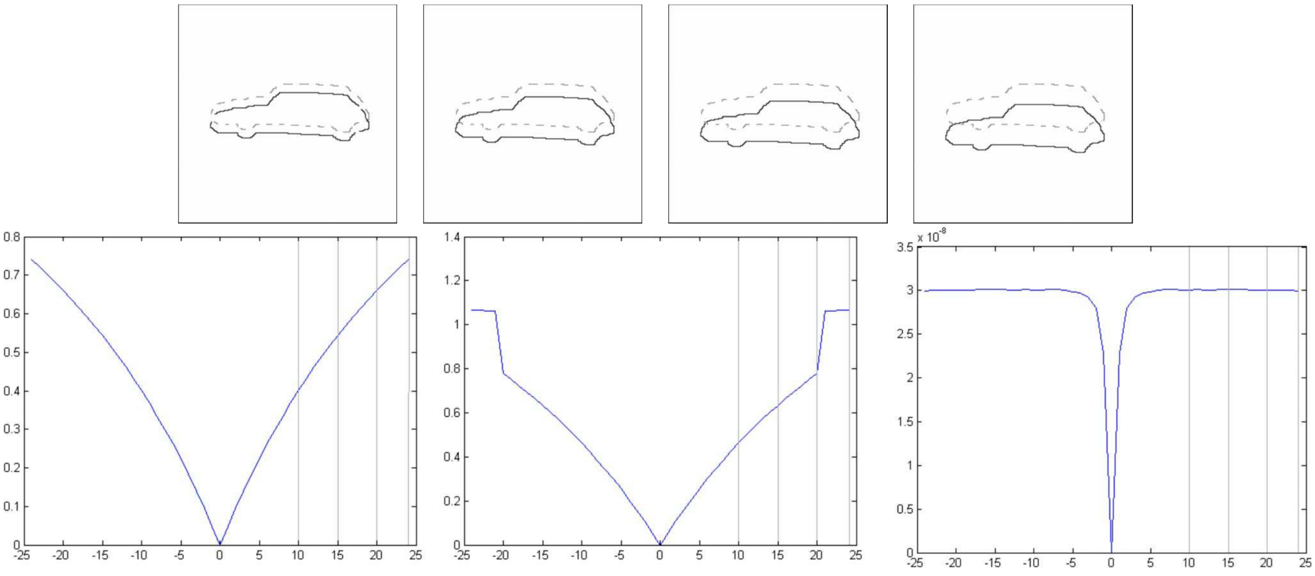
FOM metric

HAU metric

PAS metric

$HAF1$ metric

Fig. 6 Examples of localization metrics evaluation for the first ground truth and a translation alteration: the x and y axes represent the parameter of the alteration, and the z axis corresponds to the evaluation metric values.



PAS metric

$HAF1$ metric

KUL metric

Fig. 7 Examples of localization metrics behaviors: monotony and continuity illustrations.

ment using synthetic results. The simulated localization result is obtained by altering the ground truth (see Fig. 3). 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 behavior.

3.1 Properties

We defined eight properties that a localization metric should fulfill. The first ones check if the localization metric M is a distance. To be a distance, a metric should fulfill the following properties:

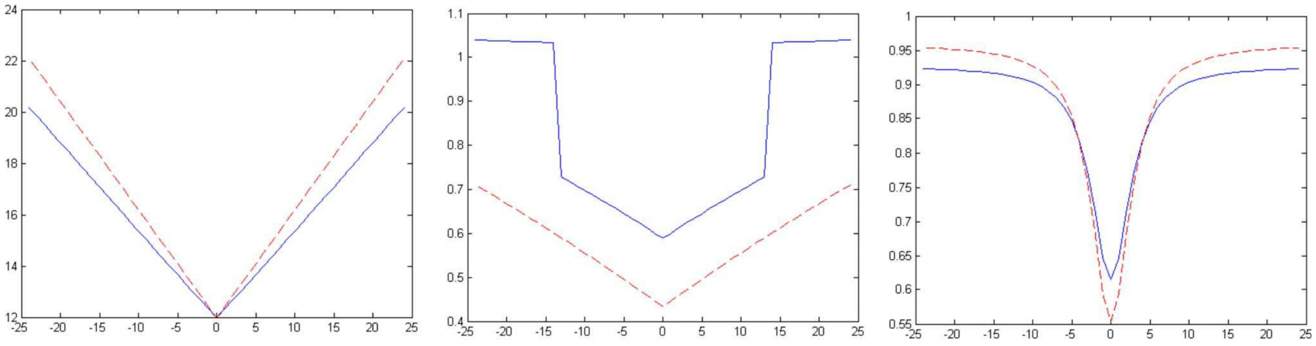
- symmetry: $M(I_1,I_2)=M(I_2,I_1)$
- separation: $M(I_1,I_2)=0 \Leftrightarrow I_1=I_2$
- triangle inequality: $M(I_1,I_3) \leqslant M(I_1,I_2)+M(I_2,I_3)$

where I_1, I_2, and I_3 are localization results.

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

1. Axial symmetry: a metric should equally penalize two results with the same alteration, but in opposite directions (for example, translations of the bounding box +5 or −5 pixels horizontally).
2. Strict monotony: a metric should penalize the results the more they are altered.
3. Uniform continuity: a metric should not have an important gap between two close results.
4. Scale dependency: a metric result should depend on the scale of the localized object.
5. Shape dependency: a metric result should depend on the shape of the localized object.

Since we consider in this work a metric as a black box providing a dissimilarity measure between a ground truth and a localization result, we need to verify experimentally if it fulfills these properties. We want this analysis of the behavior of a metric to be automatic. We propose to use an



$DMoy$ metric

$HAF1$ metric

FOM metric

Fig. 8 Examples of localization metrics behaviors: scale dependency illustrations.

Table 3 Results for distance properties checking.

Metric	Representation	Symmetry	Separation	Triangle inequality	Score
ROB_{loc}	Box				
ROB_{com}	Box	✓			*
ROB_{cor}	Box	✓			*
ErrLoc	Contour	✓	✓	✓	***
ErrSous	Contour		✓	✓	**
ErrSur	Contour		✓	✓	**
SNR	Contour				
rms	Contour	✓	✓	✓	***
Lq, 1	Contour	✓	✓	✓	***
Lq, 3	Contour	✓	✓	✓	***
KUL	Contour		✓	✓	**
BAH	Contour	✓			*
JEN	Contour		✓	✓	**
DMoy	Contour		✓	✓	**
DMoC	Contour		✓	✓	**
FOM	Contour		✓	✓	**
HAU	Contour	✓	✓	✓	***
BAD, 1	Contour	✓	✓	✓	***
BAD, 2	Contour	✓	✓	✓	***
BAD, 3	Contour	✓	✓	✓	***
ODI_n, 1	Contour		✓	✓	**
ODI_n, 2	Contour		✓	✓	**
UDI_n, 1	Contour		✓	✓	**
UDI_n, 2	Contour		✓	✓	**
PAS	Mask	✓	✓	✓	***
HEN1	Mask				
HEN2	Mask		✓	✓	**
YAS1	Mask				
YAS2	Mask				
YAS3	Mask				
MAR_{gce}	Mask	✓	✓	✓	***
MAR_{lce}	Mask	✓	✓	✓	***
HAM	Mask	✓	✓	✓	***

Table 3 (*Continued.*)

Metric	Representation	Symmetry	Separation	Triangle inequality	Score
HAF1	Mask	✓	✓	✓	***
HAF2	Mask		✓	✓	**
VIN	Mask	✓	✓	✓	***
P_{px}	Mask				
R_{px}	Mask				

experimental protocol involving a significant dataset composed of various ground truths and several localization results.

3.2 *Creation of the Localization Results Dataset*

To verify the previously mentioned properties, we need a large amount of image couples corresponding to the ground truth and the simulated localization result. To create this database, we considered 16 ground truths that can be seen in Fig. 4. We used eight ground truths representing a bounding box with different sizes and shapes. We also considered the case where a ground truth is near the border of the image. We chose to create half the database with bounding boxes, as they are widely used as a recognition result despite their poor information face for more complex shapes. We also created eight other ground truths corresponding to real objects that are traditionally used in image interpretation: a bike, man, car, etc. Those images are composed of 256 by 256 pixels. These ground truths are associated with real images such as those coming from the PASCAL VOC Challenge database. Each ground truth was available in the corresponding types of localization results: bounding box, contour, or mask.

To verify if the metrics have the required properties, we used different alterations to create synthetic localization results simulating real ones. We used four alterations: translation, scale change, 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 describes a translation along the horizontal axis. Both parameters evolve between −24 and +24 pixels, which lead to 2.400 simulated localization results for one ground truth. The scale alteration depends on two parameters as well. The parameter x denotes a scale change along the vertical axis and the y parameter denotes a scale change along the horizontal axis. Those parameters evolve also between −24 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 deg (with a progress step of 1 deg). 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 corresponds to a perspective alteration along the horizontal axis. A positive value of x corresponds to an upscale of the top of the image and a downscale of the bottom of the image. Those parameters evolve also between −24 and +24 pixels. We obtain 2.400 simulated localization results per ground truth. We can see in Fig. 5 examples of the alterations. We finally obtain a total of 118.080 synthetic localization images. Combinations of these alterations are not considered, because it would be very difficult to analyze the following behavior.

3.3 *Checking Properties*

For computation of the previously mentioned properties, we consider $M(I_{gt},I_{\text{alt}\_X\_Y})$ the result of the metric M for the alteration alt used with the parameters X and Y. We obtain a 3-D curve as shown in Fig. 6 for each triplet (metric, alteration, ground truth). We obtain a total of 2.304 evaluation results. For each one, we compute the properties listed in Sec. 3.1.

We first verify if the metric is a distance. We do this by checking the three properties of a distance. To verify if the metric is symmetric, we check if $M(I_{gt},I_{\text{alt}\_X\_Y}) = M(I_{\text{alt}\_X\_Y},I_{gt})$. We then check if the metric has the separation property: we first verify if $M(I_{gt},I_{gt})=0$ and then we verify that, for all alterations, the only 0 is for the case where $I_{\text{alt}\_X\_Y}=I_{gt}$. Finally, we check if the metric satisfies the triangle inequality, that is to say, we verify that $M(I_{gt},I_{\text{alt}\_X2\_Y2}) \leqslant M(I_{gt},I_{\text{alt}\_X1\_Y1}) + M(I_{\text{alt}\_X1\_Y1},I_{\text{alt}\_X2\_Y2})$. As this is highly time consuming, we only verify this property for the translation alteration.

To check if the metric is symmetric, we check, for a given alteration, if the result is the same for the opposite alteration. That is to say, for translation and perspective alterations, we check if $M(I_{gt},I_{\text{alt}\_X\_Y}) = M(I_{gt},I_{\text{alt}\_-X\_-Y})$. In the case of a rotation alteration, we check if $M(I_{gt},I_{\text{alt}\_D}) = M(I_{gt},I_{\text{alt}\_-D})$.

For the monotonous property, we consider the cases where X and Y are positive. We can consider the cases where X and Y are negative by adapting the criterion. The monotonous property means that the more we alter the synthetic alteration result, the more the metric must penalize

Table 4 Results for translation alteration.

Metric	Representation	Axial symmetry	Strict monotony	Uniform continuity	Scale dependency	Shape dependency	Score
ROB_{loc}	Box	✓	✓	✓	✓+	✓	*****
ROB_{com}	Box						
ROB_{cor}	Box						
ErrLoc	Contour	✓			✓−	✓	***
ErrSous	Contour	✓			✓+/−	✓	**
ErrSur	Contour	✓			✓−	✓	***
SNR	Contour	✓			✓+	✓	***
rms	Contour	✓			✓−	✓	***
Lq, 1	Contour	✓			✓−	✓	***
Lq, 3	Contour	✓			✓−	✓	***
KUL	Contour	✓			✓−	✓	***
BAH	Contour	✓			✓−	✓	***
JEN	Contour	✓			✓−	✓	***
DMoy	Contour	✓	✓	✓	✓−	✓	*****
DMoC	Contour	✓	✓	✓	✓−	✓	*****
FOM	Contour	✓	✓	✓	✓+/−	✓	****
HAU	Contour	✓	✓	✓			***
BAD, 1	Contour	✓		✓	✓−	✓	****
BAD, 2	Contour	✓			✓+	✓	***
BAD, 3	Contour	✓			✓+	✓	***
ODI_n, 1	Contour	✓			✓+/−	✓	**
ODI_n, 2	Contour	✓			✓−	✓	***
UDI_n, 1	Contour	✓			✓+/−	✓	**
UDI_n, 2	Contour	✓			✓−	✓	***
PAS	Mask	✓	✓	✓	✓+	✓	*****
HEN1	Mask						
HEN2	Mask	✓	✓	✓	✓+	✓	*****
YAS1	Mask	✓	✓	✓	✓+	✓	*****
YAS2	Mask	✓	✓	✓	✓−	✓	*****
YAS3	Mask	✓	✓	✓	✓−	✓	*****
MAR_{gce}	Mask	✓	✓	✓	✓−	✓	*****
MAR_{lce}	Mask	✓	✓	✓	✓−	✓	*****
HAM	Mask	✓		✓	✓−	✓	****

Table 4 (*Continued.*)

Metric	Representation	Axial symmetry	Strict monotony	Uniform continuity	Scale dependency	Shape dependency	Score
HAF1	Mask	✓	✓		✓+	✓	****
HAF2	Mask	✓	✓		✓−	✓	****
VIN	Mask	✓	✓	✓	✓−	✓	*****
P_{px}	Mask	✓	✓	✓	✓+	✓	*****
R_{px}	Mask	✓	✓	✓	✓+	✓	*****

the alteration. That is to say, we check if $M(I_{gt},I_{\text{alt}\_X\_Y}) \leq M(I_{gt},I_{\text{alt}\_X+1\_Y})$ or $M(I_{gt},I_{\text{alt}\_X\_Y}) \leq M(I_{gt},I_{\text{alt}\_X\_Y+1})$. We also check if $M(I_{gt},I_{\text{alt}\_X\_Y}) \leq M(I_{gt},I_{\text{alt}\_X+1\_Y+1})$. For the rotation alteration, the criterion is easier: $M(I_{gt},I_{\text{alt}\_D}) \leq M(I_{gt},I_{\text{alt}\_D+1})$, but we compute it only for alterations between 0 and 45 deg. We can adapt these criteria for the strictly monotonous property by using strict inequality.

For the continuity property, we also considered the cases where X and Y are positive. The metric is continuous if the difference between two consecutive results (that is to say, $M(I_{gt},I_{\text{alt}\_X\_Y})$ and $M(I_{gt},I_{\text{alt}\_X+1\_Y})$ for example) is not too high regarding the whole amplitude of the metric. So, we set a threshold to be $1/8$ of the amplitude of the metric, and we check if $|M(I_{gt},I_{\text{alt}\_X\_Y})-M(I_{gt},I_{\text{alt}\_X+1\_Y})| <= T$. The same criterion is used for the Y parameter. If the difference is higher than the threshold, there are two possibilities: there is a noncontinuation or the slope is very high. To clear the ambiguity, we compare this difference with the previous and next ones, and we consider that the metric is not continuous if it is four time higher, that is to say if $|M(I_{gt},I_{\text{alt}\_X\_Y})-M(I_{gt},I_{\text{alt}\_X+1\_Y})| > 4*|M(I_{gt},I_{\text{alt}\_X-1\_Y}) - M(I_{gt},I_{\text{alt}\_X\_Y})|$, and $|M(I_{gt},I_{\text{alt}\_X\_Y})-M(I_{gt},I_{\text{alt}\_X+1\_Y})| > 4*|M(I_{gt},I_{\text{alt}\_X+1\_Y})-M(I_{gt},I_{\text{alt}\_X+2\_Y})|$.

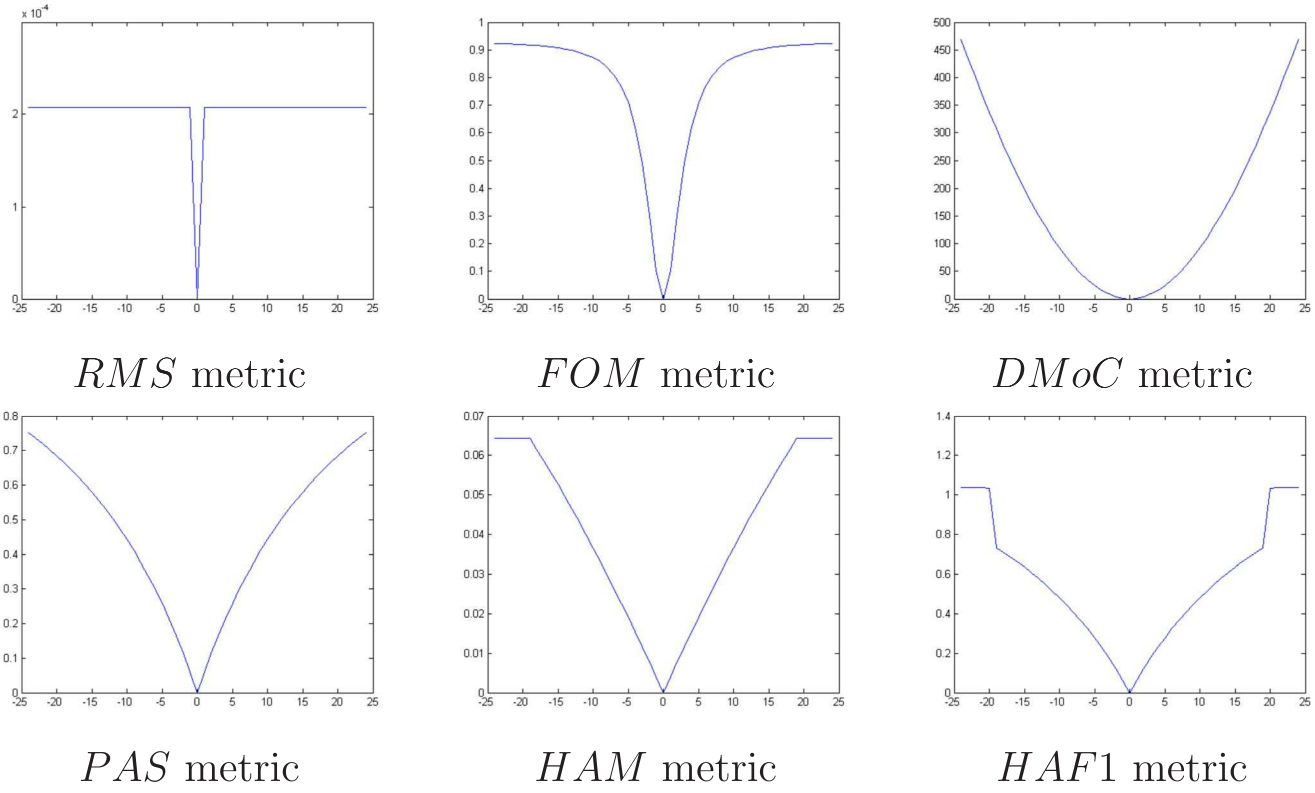
We can see in Fig. 7 four cases of alteration of the same ground truth representing a car. This alteration consists in a translation along the vertical axis of 10, 15, 20, and 24 pixels. The three curves present the evaluation results of three different metrics (PAS, HAF1, and KUL metrics) for the same localization results, i.e., the same ground truth and alteration. The curves represent the value of the metrics for different values of translation along the vertical axis. First, we can see that all metrics are symmetric. The first metric is strictly monotonous and continuous; the result is increasing with the translation, and there is no gap. On the contrary, we can see on the second metric that there is a gap when the car is translated from 20 pixels, so the metric is not continuous but is still strictly monotonous. The third metric has a huge slope but no gap, so it is continuous, and we can see that the metric equally penalizes alterations of 10, 15, 20, and 24 pixels, so the metric is not strictly monotonous.

Finally, we are looking for the dependency of the result to the shape or the scale of the localized object. To check if the shape is important, we compare results from the fourth and the fifth ground truths. These two ground truths represent both a rectangle, with the same area and the same circumference, but one is horizontal while the other is vertical. If the results from these two ground truths are similar, the metric does not take into account the shape of the localized object. Concerning the dependency to the scale, we compare results from the first and the second ground truths, which both represent a square but with different scales. We can distinguish four cases.

- Results can be independent of the scale of the localized object.
- Results can be dependent in three different ways:

– a metric can penalize a smaller object less (noted $\surd-$), because it is smaller, it is less relevant
– a metric can penalize a smaller object more (noted $\surd+$), because it is smaller, the alteration is more relevant
– a metric can penalize a smaller object more for some alterations and less for some others (noted $\surd+/-$).

The expected good results are the cases where a metric can, at the same time, take into account the size of the original object and be coherent: only the cases noted $\surd-$ and $\surd+$ are then satisfactory. We do not put forward these cases, as both can be correct regarding the application. The case noted $\surd+/-$ is clearly not expected to be correct. We do not consider correct the case where the metric's result is independent of the object, since it cannot bring information during the evaluation process. We can see in Fig. 8 the three different cases where the evaluation result is dependent on the size of the localized object. The considered alteration is a translation along the vertical axis, with a constant alteration along the horizontal axis of 24 pixels. The plain curve corresponds to the evaluation result for the first ground truth, and the dashed curve corresponds to the evaluation for the second ground truth, which is the biggest one. We can see that the Moy metric penalizes a bigger object more, whereas the HAF1 metric penalizes a smaller object more. We can also notice that the noncontinuity of the HAF1 metric appears with the smaller object, but not with the big one. The FOM metric penalizes the smaller object more for a small alteration, and penalizes a bigger object more for a larger alteration.



RMS metric

FOM metric

$DMoC$ metric

PAS metric

HAM metric

$HAF1$ metric

Fig. 9 Some evaluation results for the translation alteration.

4 Experimental Results

In this section, we present the comparative study of 33 evaluation metrics and discuss the obtained results.

4.1 *Distance*

The results in Table 3 show which metrics satisfy the properties of a distance. We can see that most of the metrics do not satisfy the three properties, particularly the symmetric one. The symmetric property is interesting, but if a metric does not have it, it can be easily managed by taking it into account during the evaluation process. The most interesting property is the separation property, as it enables us to see that one metric does not penalize some alterations at all. We can also notice that when a metric satisfies the separation property, it also satisfies the triangle inequality. One should notice that the three properties are needed for a metric to be a distance. These results enable us to see which metrics are not distances by the presence of at least one exception.

4.2 *Translation*

Table 4 shows the results obtained for the translation alteration. First, we can see that most metrics perform quite well, with at least three criteria out of five. Three metrics do not penalize translation at all: $\mathrm{ROB}_{\mathrm{com}}$, $\mathrm{ROB}_{\mathrm{cor}}$, and HEN1. This can be easily explained for the two metrics from the Robin project, as each metric is dedicated to one type of alteration: $\mathrm{ROB}_{\mathrm{loc}}$ is dedicated to penalize the translation of the center of the bounding box, whereas $\mathrm{ROB}_{\mathrm{com}}$ and $\mathrm{ROB}_{\mathrm{cor}}$ penalize the shape of the bounding box. As HEN1 only takes into account information about pixel numbers, it is insensible to translation. We can also notice that the HAU metric is the only one that does not take the size or the shape of the localized object into account. We can see that all contour-based metrics, except DMoy, DMoC, FOM, and HAU, give worse results than mask-based metrics, mainly because they do not fulfill the strict monotony and the uniform continuity properties. The region-based metrics fulfill at least four properties out of five.

We can see in Fig. 9 some evaluation results for different metrics, with a translation along the diagonal, that is to say with the alt_X_X alteration on the first ground truth. We can see that the rms metric judges equally a translation of 1 or several pixels. The metrics ErrLoc, ErrSous, ErrSur, SNR, Lq, KUL, BAH, and JEN behave the same way. The figure of merit of Pratt FOM tends to behave similarly but is less restrictive, so it is monotonous and continuous. The PAS metric tends to behave linearly, whereas the DMoC metric highly penalizes only huge translations. The Odet and Baddeley metrics tend to behave like the PAS or the DMoC metrics, but present a noncontinuity and are not strictly monotonous. The HAM metric shows that it is not strictly monotonous after translation of 19 pixels. The HAF1 metric behaves like the PAS metric, but can be noncontinuous if the translation is too high. Concerning the other region-based metrics, they behave like PAS.

4.3 *Scale*

We can see in Table 5 the results obtained for a scale alteration. The three metrics from the Robin project behave as expected: the $\mathrm{ROB}_{\mathrm{loc}}$ metric does not penalize scaling, because the center of the object does not change, whereas the two other metrics $\mathrm{ROB}_{\mathrm{com}}$ and $\mathrm{ROB}_{\mathrm{cor}}$ penalize it. We can

Table 5 Results for scale alteration.

Metric	Representation	Strict monotony	Uniform continuity	Scale dependency	Shape dependency	Score
ROB_{loc}	Box					
ROB_{com}	Box	✓	✓	✓+		***
ROB_{cor}	Box		✓	✓+	✓	***
ErrLoc	Contour			✓−	✓	**
ErrSous	Contour			✓+/−	✓	*
ErrSur	Contour			✓−	✓	**
SNR	Contour			✓+	✓	**
rms	Contour			✓−	✓	**
Lq, 1	Contour			✓−	✓	**
Lq, 3	Contour			✓−	✓	**
KUL	Contour			✓−	✓	**
BAH	Contour	✓	✓	✓−	✓	****
JEN	Contour			✓−	✓	**
DMoy	Contour	✓	✓	✓+/−	✓	***
DMoC	Contour	✓	✓	✓+/−	✓	***
FOM	Contour	✓	✓	✓+/−	✓	***
HAU	Contour	✓	✓			**
BAD, 1	Contour		✓	✓+/−	✓	**
BAD, 2	Contour		✓	✓+	✓	***
BAD, 3	Contour		✓	✓+	✓	***
ODI_n, 1	Contour			✓+/−	✓	*
ODI_n, 2	Contour			✓+/−	✓	*
UDI_n, 1	Contour			✓+/−	✓	*
UDI_n, 2	Contour			✓+/−	✓	*
PAS	Mask	✓	✓	✓+	✓	****
HEN1	Mask	✓	✓	✓+/−	✓	***
HEN2	Mask	✓	✓	✓+	✓	****
YAS1	Mask		✓	✓+	✓	***
YAS2	Mask		✓	✓−	✓	***
YAS3	Mask		✓	✓−	✓	***
MAR_{gce}	Mask	✓	✓	✓−	✓	****
MAR_{lce}	Mask	✓	✓	✓−	✓	****
HAM	Mask		✓	✓−	✓	***

Table 5 (*Continued.*)

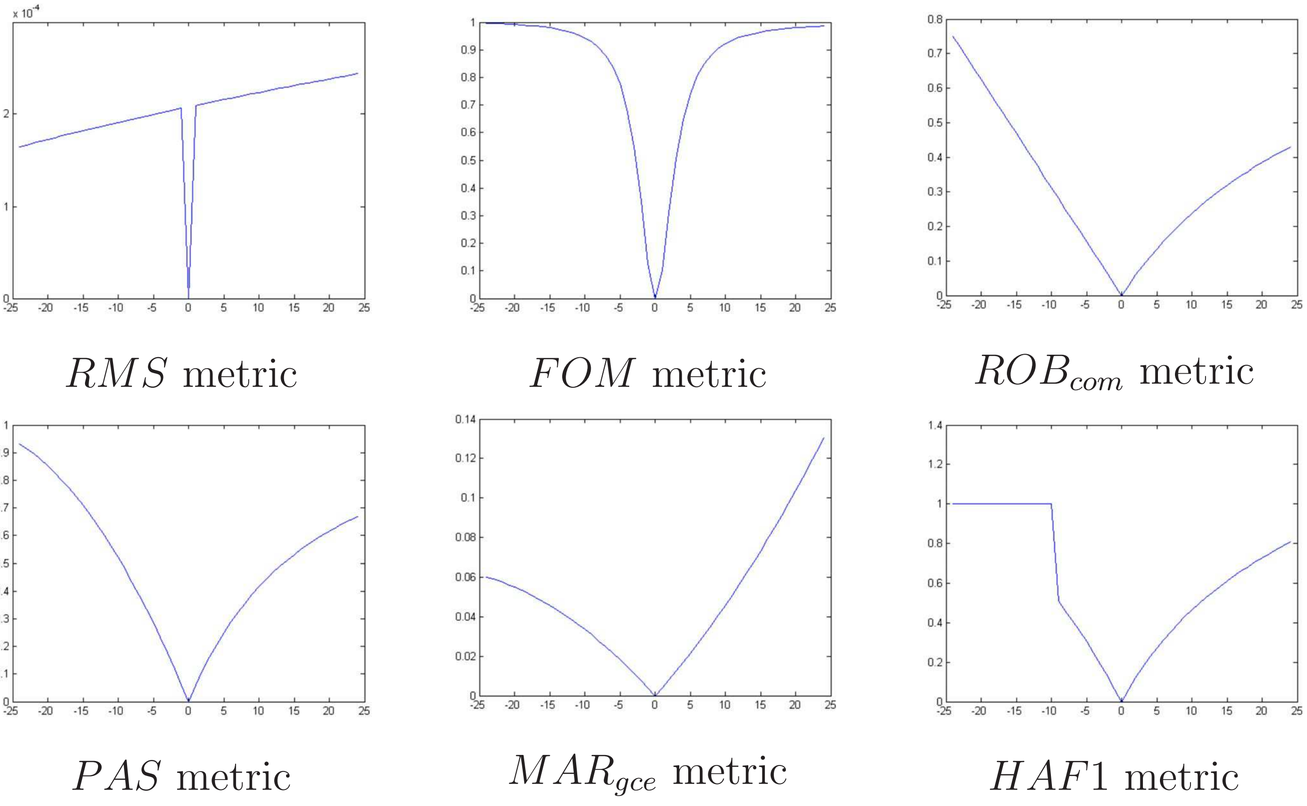
Metric	Representation	Strict monotony	Uniform continuity	Scale dependency	Shape dependency	Score
HAF1	Mask			✓+/−	✓	*
HAF2	Mask	✓		✓−	✓	***
VIN	Mask	✓	✓	✓−	✓	****
P_{px}	Mask		✓	✓+	✓	***
R_{px}	Mask		✓	✓+	✓	***

also notice that only the HAU metric is not dependent on size or shape of the localized object for the scale alteration as well.

We can see in Fig. 10 some evaluation results for different metrics with a homothetic alteration, that is to say, with an alt_X_X alteration, on the first ground truth. A problem can be seen with the rms metric, as it penalizes more a downscaling of 1 pixel than a downscaling of 20 pixels. The ErrLoc, ErrSous, ErrSur, SNR, Lq, BAH, KUL, and JEN metrics have the same error, which is also present considering other ground truths. We can see that the figure of merit FOM penalizes almost equally downscaling and upscaling. The DMoy, DMoC, HAU, BAD, 1, and Odet metrics tend to penalize almost equally downscaling and upscaling like the FOM metric. The PAS and ROB_{com} metrics penalize much more a downscaling, whereas the MAR_{gce} metric penalizes more upscaling. The YAS2 and HAF1 metrics penalize more downscaling like the PAS metric. The Henricsson, YAS1, YAS3, MAR_{lce}, HAM, HAF2, and IN metrics penalize more upscaling like the MAR_{gce} metric. We can notice that the HAF1 metric still presents a noncontinuity, but only for a downscaling, whereas the HAF2 presents a noncontinuity for upscaling.

4.4 *Rotation*

We can see in Table 6 results obtained for the rotation alteration. We can notice that two metrics are not able to penalize this kind of alteration: ROB_{loc} and HEN1. We can also notice that all metrics are symmetric, that is to say,



RMS metric

FOM metric

ROB_{com} metric

PAS metric

MAR_{gce} metric

$HAF1$ metric

Fig. 10 Some evaluation results for the scale alteration.

Table 6 Results for rotation alteration.

Metric	Representation	Axial symmetry	Strict monotony	Uniform continuity	Scale dependency	Shape dependency	Score
$\mathrm{ROB}_{\mathrm{loc}}$	Box						
$\mathrm{ROB}_{\mathrm{com}}$	Box	✓		✓	✓+/−		**
$\mathrm{ROB}_{\mathrm{cor}}$	Box	✓		✓		✓	***
ErrLoc	Contour	✓		✓	✓−		***
ErrSous	Contour	✓		✓	✓+/−		**
ErrSur	Contour	✓		✓	✓−		***
SNR	Contour	✓			✓+		**
rms	Contour	✓		✓	✓−	✓	****
Lq, 1	Contour	✓		✓	✓−	✓	****
Lq, 3	Contour	✓		✓	✓−	✓	****
KUL	Contour	✓		✓	✓−	✓	****
BAH	Contour	✓			✓+/−	✓	**
JEN	Contour	✓		✓	✓−	✓	****
DMoy	Contour	✓	✓	✓	✓−		****
DMoC	Contour	✓	✓	✓	✓−		****
FOM	Contour	✓	✓	✓	✓−	✓	*****
HAU	Contour	✓		✓	✓−		***
BAD, 1	Contour	✓	✓	✓	✓−	✓	*****
BAD, 2	Contour	✓	✓	✓	✓−		****
BAD, 3	Contour	✓	✓	✓	✓+/−	✓	****
ODI_n, 1	Contour	✓	✓	✓	✓−	✓	*****
ODI_n, 2	Contour	✓	✓	✓	✓−	✓	*****
UDI_n, 1	Contour	✓	✓	✓	✓−	✓	*****
UDI_n, 2	Contour	✓	✓	✓	✓−	✓	*****
PAS	Mask	✓	✓	✓	✓+/−		***
HEN1	Mask						
HEN 2	Mask	✓	✓	✓	✓+/−		***
YAS1	Mask	✓	✓	✓	✓+/−		***
YAS2	Mask	✓	✓	✓	✓−		****
YAS3	Mask	✓	✓	✓	✓−		****
$\mathrm{MAR}_{\mathrm{gce}}$	Mask	✓	✓	✓	✓−	✓	*****
$\mathrm{MAR}_{\mathrm{lce}}$	Mask	✓	✓	✓	✓−	✓	*****
HAM	Mask	✓	✓	✓	✓−		****

Table 6 (*Continued.*)

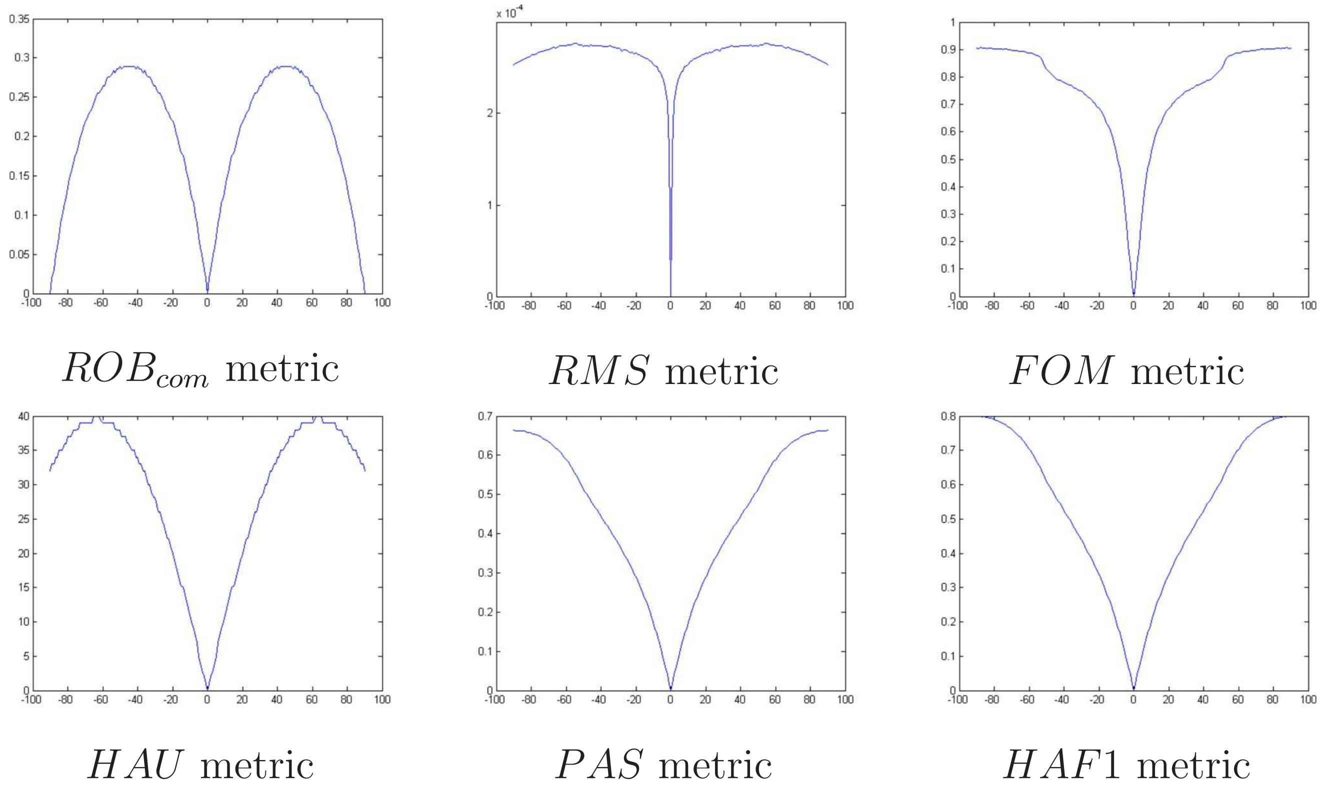
Metric	Representation	Axial symmetry	Strict monotony	Uniform continuity	Scale dependency	Shape dependency	Score
HAF1	Mask	✓	✓	✓	✓−		****
HAF2	Mask	✓	✓	✓	✓−		****
VIN	Mask	✓	✓	✓	✓−		****
P_{px}	Mask	✓	✓	✓	✓+/−		***
R_{px}	Mask	✓	✓	✓	✓+/−		***

they penalize equally a rotation of D or $-D$ deg. We can see that mask-based metrics perform globally better than contour-based ones, with at least three properties out of five.

We can see in Fig. 11 some evaluation results for a rotation alteration on the fourth ground truth. We can see that all metrics increase until an alteration of 45 deg, but there are different behaviors if the rotation exceeds 45 deg. The ROB_{com} metric is then decreasing until it is not penalizing the alteration at all. The BAH metric behaves the same way. The rms metric and the HAU distance have the same behavior but the decrease is much more limited, just like ErrLoc, ErrSous, ErrSur, SNR, Lq, KUL, and JEN metrics. Other metrics do not decrease after 45 deg and continue on correctly penalizing the alteration.

4.5 *Perspective*

We can see in Table 7 results obtained for the perspective alteration. We can notice that the two same metrics are not able to penalize this kind of alteration: ROB_{loc} and HEN1. Moreover, the BAH is also not able to correctly penalize this alteration. We can notice that all region-based metrics, except the HEN1 metric, obtain the maximal score of 5. In a general way, contour-based metrics are not able to perform as well as region-based ones.



ROB_{com} metric

RMS metric

FOM metric

HAU metric

PAS metric

$HAF1$ metric

Fig. 11 Some evaluation results for the rotation alteration.

Table 7 Results for perspective alteration.

Metric	Representation	Axial symmetry	Strict monotony	Uniform continuity	Scale dependency	Shape dependency	Score
ROB_{loc}	Box						
ROB_{com}	Box				✓+/−	✓	*
ROB_{cor}	Box	✓		✓	✓+	✓	****
ErrLoc	Contour	✓		✓	✓−	✓	****
ErrSous	Contour	✓		✓	✓+/−	✓	***
ErrSur	Contour			✓	✓−	✓	***
SNR	Contour	✓			✓+	✓	***
rms	Contour	✓			✓−	✓	***
Lq, 1	Contour	✓		✓	✓−	✓	****
Lq, 3	Contour	✓		✓	✓−	✓	***
KUL	Contour	✓		✓	✓−	✓	****
BAH	Contour						
JEN	Contour	✓		✓	✓−	✓	****
DMoy	Contour			✓	✓+/−	✓	**
DMoC	Contour			✓	✓+/−	✓	**
FOM	Contour	✓		✓	✓+/−	✓	***
HAU	Contour	✓	✓	✓			***
BAD, 1	Contour			✓	✓+/−	✓	**
BAD, 2	Contour			✓	✓+	✓	***
BAD, 3	Contour			✓	✓+	✓	***
ODI_n, 1	Contour			✓	✓+/−	✓	**
ODI_n, 2	Contour			✓	✓+/−	✓	**
UDI_n, 1	Contour	✓		✓	✓+/−	✓	***
UDI_n, 2	Contour			✓	✓+/−	✓	**
PAS	Mask	✓	✓	✓	✓+	✓	*****
HEN1	Mask						
HEN 2	Mask	✓	✓	✓	✓+	✓	*****
YAS1	Mask	✓	✓	✓	✓+	✓	*****
YAS2	Mask	✓	✓	✓	✓−	✓	*****
YAS3	Mask	✓	✓	✓	✓−	✓	*****
MAR_{gce}	Mask	✓	✓	✓	✓−	✓	*****
MAR_{lce}	Mask	✓	✓	✓	✓−	✓	*****
HAM	Mask	✓	✓	✓	✓−	✓	*****

Table 7 (*Continued.*)

Metric	Representation	Axial symmetry	Strict monotony	Uniform continuity	Scale dependency	Shape dependency	Score
HAF1	Mask	✓	✓	✓	✓+	✓	*****
HAF2	Mask	✓	✓	✓	✓−	✓	*****
VIN	Mask	✓	✓	✓	✓−	✓	*****
P_{px}	Mask	✓	✓	✓	✓+	✓	*****
R_{px}	Mask	✓	✓	✓	✓+	✓	*****

4.6 *Discussion*

Results gathering distance property checking and global behavior faces for each alteration are summarized in Table 8. We also sum each score to obtain a final score for each metric. The maximum score is 22.

We first can see that the three metrics using bounding boxes do not perform very well, with scores between 5 and 11. However, we should notice that these three metrics have been thought to be used together to obtain a final decision concerning the quality of a localization result. As each metric is able to penalize a particular kind of alteration, used together these metrics should enable a correct evaluation. Second, we can see that the metrics using contour-based representation of localization results generally give quite bad results compared to region-based ones. Except for the HEN1 metric, the minimal score for region-based metrics is 16, whereas only the DMoy, FOM, and BAD metrics have a score of 16 or higher. We can conclude that the region-based representation of localization results should be preferred. Moreover, we advise using the Martin metrics MAR_{gce} and MAR_{lce} to evaluate an interpretation algorithm.

5 Illustrations

In this section, we present some evaluation results to illustrate the conclusions set out during the comparative study. In Fig. 12, we present an original image representing a plane and the associated ground truth (from the PASCAL VOC Challenge dataset). Three localization results are also presented and have been obtained by a reference active contour segmentation algorithm[16] after different numbers of iterations. As the image is quite easy to analyze, the active contour process works well. We can consider that the obtained results correspond to an upscaling alteration. The four considered metrics behave correctly with upscaling regarding Table 5. We can see that the metrics work quite well, as they all consider the last segmentation as the best one because it is the case visually.

If we consider a more complicated case such as the one presented in Fig. 13, we can see that too many iterations in the active contour process give a bad result. Once again, we can consider that the first result corresponds to an upscaling alteration. With more iterations, we can compare the two obtained results as corresponding to a downscaling combined with a translation. We can see that the BAH metric is not able to correctly penalize this problem. We already identified some problems in its behavior in the comparative study. In Tables 4 and 5, the BAH metric obtains a lower score than FOM, PAS, and MAR_{gce} metrics.

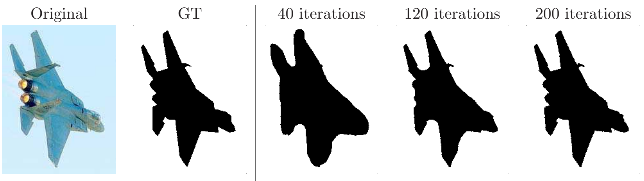
Finally, we can see in Fig. 14 another complicated case. The first result corresponds to an upscale alteration. The second one is quite correct, but the third one is quite altered. We can see that with too many iterations, the cheek of the lady is considered background. It corresponds to a partial lack of localization, and thus does not correspond to an alteration considered in the comparative study. The analysis of this kind of alteration would have necessitated too many computations (localization of the alteration within its contour and importance of the alteration). We can see that only the MAR_{gce} metric is able to penalize this problem.

6 Conclusions

We propose an evaluation protocol that permits one to verify some properties that a localization metric should fulfill. We mainly see that region-based metrics perform better than contour-based ones. That is why we advise using a region-based representation for localization algorithms. Metrics such as PAS, VIN, or both MAR_{lce} and MAR_{gce} would enable a good evaluation of localization algorithms. However, in the case where a contour-based representation of a localization result is available, metrics such as the figure of merit of Pratt FOM, the mean distance DMoy, and Baddeley distance BAD could be used. We also notice that the Baddeley distance is parameterizable, so it can be adjusted to a specific application. We can finally recommend using region-based metrics, like MAR_{lce} or MAR_{gce}, to evaluate an algorithm using bounding boxes as representation of localization results. It would permit a better evaluation than Robin metrics, even if they are dedicated to bounding boxes. If the use of region-based metrics in the case of a bounding box is not possible (for time or memory complexity), we, recommend to jointly use the three Robin

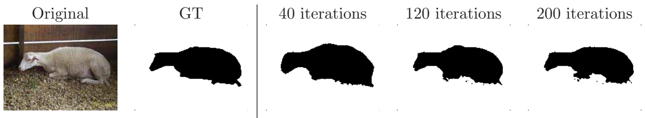
Table 8 Synthesis of obtained results.

Metric	Representation	Distance	Translation	Scale	Rotation	Perspective	Final score
ROB$_{loc}$	Box		*****				5
ROB$_{com}$	Box	*		***	**	*	7
ROB$_{cor}$	Box	*		***	***	****	11
ErrLoc	Contour	***	***	**	***	****	15
ErrSous	Contour	**	**	*	**	***	10
ErrSur	Contour	**	***	**	***	***	13
SNR	Contour		***	**	**	***	10
RMS	Contour	***	***	**	****	***	15
Lq, 1	Contour	***	***	**	****	****	16
Lq, 3	Contour	***	***	**	***	***	14
KUL	Contour	**	***	**	****	****	15
BAH	Contour	*	***	****	**		10
JEN	Contour	**	***	**	****	****	15
DMoy	Contour	**	*****	***	****	**	16
DMoC	Contour	**	*****	***	****	**	16
FOM	Contour	**	****	***	*****	***	17
HAU	Contour	***	***	**	***	***	14
BAD, 1	Contour	***	****	**	*****	**	16
BAD, 2	Contour	***	***	***	****	***	16
BAD, 3	Contour	***	***	***	****	***	16
ODI$_n$, 1	Contour	**	**	*	*****	**	12
ODI$_n$, 2	Contour	**	***	*	*****	**	13
UDI$_n$, 1	Contour	**	**	*	*****	***	13
UDI$_n$, 2	Contour	**	***	*	*****	**	13
PAS	Mask	***	*****	****	***	*****	20
HEN1	Mask			***			3
HEN 2	Mask	**	*****	****	***	*****	19
YAS1	Mask		*****	***	***	*****	16
YAS2	Mask		*****	***	****	*****	17
YAS3	Mask		*****	***	****	*****	17
MAR$_{gce}$	Mask	***	*****	****	*****	*****	22
MAR$_{lce}$	Mask	***	*****	****	*****	*****	22
HAM	Mask	***	****	***	****	*****	19
HAF1	Mask	***	****	*	****	*****	17
HAF2	Mask	**	****	***	****	*****	18
VIN	Mask	***	*****	****	****	*****	21
P_{px}	Mask		*****	***	***	*****	16
R_{px}	Mask		*****	***	***	*****	16



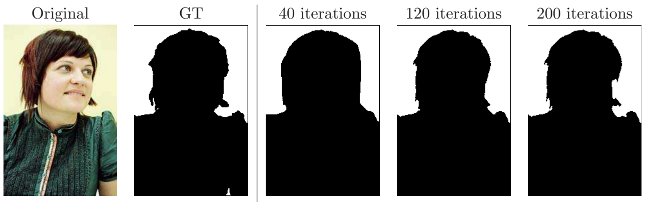
	40 iterations	120 iterations	200 iterations
BAH	0.6900	0.6892	**0.6888**
FOM	0.8905	0.3339	**0.2411**
PAS	0.7249	0.9224	**0.9416**
MAR_{gce}	0.1359	0.0305	**0.0236**

Fig. 12 Illustration 1.



	40 iterations	120 iterations	200 iterations
BAH	0.6923	**0.6919**	**0.6919**
FOM	0.6814	**0.4203**	0.4337
PAS	0.8507	**0.9271**	0.9021
MAR_{gce}	0.0535	**0.0455**	0.0515

Fig. 13 Illustration 2.



	40 iterations	120 iterations	200 iterations
BAH	0.6919	0.6911	**0.6906**
FOM	0.7048	0.5111	**0.4695**
PAS	0.0581	0.1119	**0.1145**
MAR_{gce}	0.8443	**0.7358**	0.7554

Fig. 14 Illustration 3.

metrics, as recommended in Ref. 5. The use of these metrics independently produces bad evaluation results.

It is important to notice that chosen properties (continuity, size dependency, etc.) are intuitive. To check if these properties are well chosen, we plan to do a subjective study. Many people will be asked to compare several localization results and tell us which one corresponds the best to a given ground truth. This will enable us to see if our metrics ranking corresponds to the human one, and also see if the studied properties are well chosen.

This work opens a new area of research in the analysis of evaluation metrics for image understanding. This general methodology could be improved by considering other alterations, for example the local alteration of a contour. It could be also possible to weight the error of a localization result based on semantic information. For example, if we aim to evaluate human detection algorithms, we could take into account different parts of the body such as the head. A localization error in the neighborhood of the head could be penalized more strictly than near the hands, depending in this case.

References

1. R. Cucchiara, C. Grana, M. Piccardi, and A. Prati, "Detecting moving objects, ghosts, and shadows in video streams," *IEEE Trans. Pattern Anal. Mach. Intell.* **25**(10), 1337–1342 (2003).
2. N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," *J. Comput. Vis Pattern Recogn. (CVPR)* **1**, 886–893 (2005).
3. F. Jurie and C. Schmid, "Scale-invariant shape features for recognition of object categories," *IEEE Comput. Soc. Conf. Comput. Vis Patt. Recog. (CVPR)* **2**, 90–96 (2004).
4. M. Everingham, A. Zisserman, C. Williams, L. Van Gool, M. Allan, C. Bishop, O. Chapelle, N. Dalal, T. Deselaers, G. Dorko, *et al.*, "The 2005 pascal visual object classes challenge," see http://www.pascal-network.org/challenges/VOC/ (2005).
5. E. D'Angelo, S. Herbin, and M. Ratiéville, "Robin challenge evaluation principles and metrics," see http://robin.inrialpes.fr (2006).
6. M. Basseville, "Distance measures for signal processing and pattern recognition," *J. Signal Process.* **18**(4), 349–369 (1989).
7. D. L. Wilson, A. J. Baddeley, and R. A. Owens, "A new metric for grey-scale image comparison," *Intl. J. Comput. Vis* **24**(1), 5–17 (1997).
8. 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 IEEE Intl. Conf. Computer Vision (ICCV)* **2**, 416–423 (Jul. 2001).
9. A. Hafiane, S. Chabrier, C. Rosenberger, and H. Laurent, "A new supervised evaluation criterion for region based segmentation methods," *LNCS4678 Adv. Concepts Intell. Vision Syst. (ACIVS)* **4678**, 439–448 (2007).
10. A. J. Baddeley, "An error metric for binary images," *Robust Computer Vision*, pp. 59–78 (1992).
11. M. Beauchemin, KP. B. Thomson, and G. Edwards, "On the hausdorff distance used for the evaluation of segmentation results," *Can. J. Remote Sens.* **24**(1), 3–8 (1998).
12. B. Hemery, H. Laurent, C. Rosenberger, and B. Emile, "Evaluation protocol for localization metrics—application to a comparative study," *Intl. Conf. Image Signal Process. (ICISP)*, pp. 273–280 (2008).
13. W. Pratt, O. D. Faugeras, and A. Gagalowicz, "Visual discrimination of stochastic texture fields," *IEEE Trans. Syst. Man Cybern.* **8**(11), 796–804 (1978).
14. R. Myers, R. C. Wison, and E. R. Hancock, "Bayesian graph edit distance," *IEEE Trans. Pattern Anal. Mach. Intell.* **22**(6), 628–635 (2000).
15. T. B. Sebastian, P. N. Klein, and B. B. Kimia, "Recognition of shapes by editing shock graphs," *Intl. Conf. Computer Vision (ICCV)*, pp. 755–762 (2001).
16. T. F. Chan and L. A. Vese, "Active contours without edges," *IEEE Trans. Image Process.* **10**(2), 266–277 (2001).
17. S. Chabrier, "Contribution à l'évaluation de performances en segmentation d'images," PhD Thesis, Univ. d'Orléans, Bourges, France (2005).
18. D. Coquin, P. Bolon, and Y. Chehadeh, "Evaluation quantitative d'images filtrées," *Colloque GRETSI* **2**, 1351–1354 (1997).
19. T. Peli and D. Malah, "A study of edge detection algorithms," *J. Comput. Graph. Image Process.* **20**, 1–21 (1982).
20. C. Odet, B. Belaroussi, and H. Benoit-Cattin, "Scalable discrepancy measures for segmentation evaluation," *Intl. Conf. Image Process. (ICIP)*, pp. 785–788 (2002).
21. O. Henricsson and E. Baltsavias, "3-d building reconstruction with aruba: a qualitative and quantitative evaluation," *Automatic Extraction of Man-Made Objects from Aerial and Space Images (2)*, pp. 65–76, Springer, Verlag, Berlin (1997).
22. W. A. Yasnoff, J. K. Mui, and J. W. Bacus, "Error measures for scene segmentation," *J. Patt. Recog.* **9**, 217–231 (1977).
23. Q. Huang and B. Dom, "Quantitative methods of evaluating image segmentation," *Intl. Conf. Image Processing (ICIP)* **3**, 53–56 (1995).
24. A. Hafiane, "Caractrisation de textures et segmentation pour la recherche d'images par le contenu," PhD Thesis, Univ. de Paris-Sud XI (2005).
25. L. Vinet, "Segmentation et mise en correspondance de régions de paires d'images stéréoscopiques," PhD Thesis, Univ. de Paris IX Dauphine (1991).
26. J. P. Cocquerez and S. Philipp, *Analyse d'Images: Filtrage et Segmentation*, Masson (1995).
27. C. Wolf and J. M. Jolion, "Object count/area graphs for the evaluation of object detection and segmentation algorithms," *Intl. J. Document Anal. Recog.* **8**(4), 280–296 (2006).

Baptiste Hemery is an assistant professor at IUT of Saint-Lô (France). He obtained his PhD from the University of Caen Basse-Normandie in 2009. He belongs to the GREYC laboratory in the image and computer security research units. His research interests concern image interpretation evaluation and biometric systems.

Helene Laurent is an assistant professor at ENSI of Bourges (France). She obtained her PhD from the University of Nantes in 1998. She belongs to the Prisme Institute of Orleans University in the ISS research unit. Her research interests concern image processing evaluation.

Bruno Emile is an assistant professor at IUT of Châteauroux (France). He obtained his PhD from the University of Nice in 1996. He belongs to the Prisme Institute of Orleans University in the ISS research unit. His research interests concern object detection and recognition.

Christophe Rosenberger is full professor at ENSICAEN (France). He obtained his PhD from the University of Rennes I in 1999. Since 2007, he has belonged to the GREYC laboratory. His research interests concern evaluation of image processing, image understanding, and biometric systems.