Edge detection: a review of dissimilarity evaluations and a proposed normalized measure

Baptiste Magnier

Edge Detection: A Review of Dissimilarity Evaluations and a Proposed Normalized Measure

Baptiste Magnier

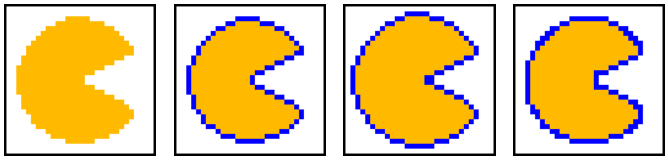
***Abstract*—In digital images, edges characterize object boundaries, so edge detection remains a crucial stage in numerous applications. To achieve this task, many edge detectors have been designed, producing different results, with various qualities of segmentation. Indeed, optimizing the response obtained by these detectors has become a crucial issue, and effective contour assessment assists performance evaluation. In this paper, several referenced-based boundary detection evaluations are detailed, pointing out their advantages and disadvantages, theoretically and through concrete examples of image edges. Then, a new normalized supervised edge map quality measure is proposed, comparing a ground truth contour image, the candidate contour image and their associated spatial nearness. The effectiveness of the proposed distance measure is demonstrated theoretically and through several experiments, comparing the results with the methods detailed in the state-of-the-art. In summary, compared to other boundary detection assessments, this new method proved to be a more reliable edge map quality measure.**

***Index Terms*—Edge detection, supervised evaluation, distance measures.**

I. Introduction

In computer science, all systems, especially automated information processing structures, must be evaluated before being developed, principally for industrial applications or medical data. Image processing is no exception to this rule. In image analysis, image segmentation is one of the most critical tasks and represents an essential step in low-level vision. All the methods developed therefore have to be tested and assessed, whether regarding edge detection, point matching, region segmentation or image restoration/enhancement.

Edge detection represents one of the pioneer theoretical works in image processing tasks [38] and remains a key point in many applications. It is extensively used because boundaries include the most important structures in the image [54][44]. Furthermore, edge detection itself could be used to qualify a region segmentation technique [32][34]. In addition, contour extraction remains a very useful preprocessing step in image segmentation, registration, reconstruction, interpretation and tracking [33]. An efficient boundary detection method should create a contour image containing edges at their correct locations with a minimum of misclassified pixels. Edge detection assessment is therefore an essential field of study, but research has still not necessarily gone deep enough regarding contour detection for digital images. Contrary to region segmentation evaluations, which may consider color attributes, contour detection assessments generally use binary images. Thus, an edge detection evaluation leads to a direct assessment of the considered binary image. Moreover, measures of boundary detection evaluation are based on the fact that a reliable edge map should characterize all the relevant structures of the image. It should also create a minimum of spurious pixels or holes (oversights). Therefore, an evaluation can be used to assess and improve an algorithm, or to optimize edge detector parameters [21], for example thresholds in the last stage of the detection.



(a) Object (b) Contour 1 (c) Contour 2 (d) Contour 3

Fig. 1. Several edge chains are available for the same object. The image in (a) is of size 30×30. Contour points are represented in blue for image (b), (c) and (d). Contour 1 represents the inner edge whereas contour 2 corresponds to the outer boundary. Contour 3 is the result of a step edge detector.

In edge detection assessment, the measurement process can be classified as using either unsupervised or supervised evaluation criteria. The first class of methods exploits only the input data image and gives a score of coherence that qualifies the result given by the algorithm [21][17][42]. The second class computes a similarity/dissimilarity measure between a segmentation result and a ground truth obtained from synthetic data or an expert judgment [9][22][31]. As edge extraction is performed for image processing tasks and computer vision, similarity measures are important within a broad flied of study, for example in image interpretation, consisting in automatically extracting or recognizing objects in an image, compared to a model. Moreover, the assessment can compute a similarity or a dissimilarity of the shapes attributes between two of binary objects [13][22][37]. Thus, the problem of supervised edge detection assessment amounts to a question of pattern recognition. Nevertheless, the edge position of an object could be interpreted in different ways, as represented in Fig. 1. Indeed, for a vertical step edge, an edge can be located either on the left, or on the right (which corresponds to an inner or an outer contour). Considering a vertical blurred edge (i.e. ramp contour [6]), the true edge is placed in the middle of the blur, but the question of the position remains the same when the size of the blur is even. In the crest lines case [37], true edges are chosen in the middle of the ridge or of the valley when the width of the ridge/valley is equal to a odd number (i.e. at the maxima -top of ridges- or minima -bottom

Ecole des Mines d'Alès, Site de Croupillac, LGI2P, 7. rue Jules Renard, 30100 ALES, France. E-mail: baptiste.magnier@mines-ales.fr

Special thanks to Adam Clark Gimming.

of valleys- [36]). Finally, for a single real or synthetic image, as several contour chains constituting the ground truth depend on several positions of the $true$ pixels, thus the combinations of the boundary pixel placements are numerous, so many chains could be created/chosen. For example, we can create a ground truth boundary map of the image (a) in Fig. 1 with as many combinations as there are pixels in Contour 1, 2 and 3. This observation shows that the evaluation must not be binary, but rather illustrates the importance of taking into account the misplaced edge distance regarding the desired contour of the ground true image. For all these reasons, as described in [44], quantitative assessments are generally lacking in proposed edge detection methods.

In image segmentation evaluation, the Structural Similarity Index ($SSIM$) estimates the visual impact of shifts in an image [57]. This measure is based on the grayscale information concerning the attributes that represent the structure of objects in the scene. $SSIM$ consists of three local comparison functions, namely luminance comparison, contrast comparison, and structure comparison between two signals excluding other remaining errors. Unlike the $PSNR$ (Peak Signal to Noise Ratio) or $RMSE$ (Root-Mean-Square Error), which are measured at the global or the image level, the $SSIM$ is computed locally by moving a 8×8 window for each pixel. The final score of an entire image corresponds to the mean of all the local scores. Applications of this image quality evaluation include automatically judging the performance of compression or restoration algorithms, concerning grayscale or color images. Even though $SSIM$ can be applied in the case of an edge detection evaluation, in the presence of too many areas without contours, the obtained score is not efficient or useful (in order to judge the quality of edge detection with the $SSIM$, it is necessary to compare with an image having detected edges situated throughout the image areas).

This work focusses on comparisons of supervised edge detection evaluations with respect to binary representation of the boundaries. As introduced above, a supervised evaluation process estimates scores between a ground truth and a candidate edge map (both binary images). These scores could be evaluated counting the number of erroneous pixels, but also through spatial distances of misplaced or undetected contours; this paper outlines the various algorithms presented in the literature. In addition, a new supervised edge map quality measure based on the distances of misplaced pixels is presented and compared to the others; the score obtained by this measure is normalized and can be interpreted for each type of contour. In order to present the effectiveness and efficiency of the proposed method, this new formula is compared with others presented and detailed in the state-of-the-art. Thus, several comparisons of edge detection assessments are carried out for different perturbations such as: addition of false positive points, creation of false negative points, over-segmentation near the edge (until total dilation of the edge), contour displacement, modification of the image size... These experiments indicate that some measures are not appropriate for the evaluation of edge detection for each type of boundary, and an ideal measure must be one that reacts as coherently possible to reality. Therefore, this paper shows the relevance of a boundary detection assessment that takes into consideration the distance of both the false positive and the false negative points created or missed by the boundary detection process.

The remainder of this paper is organized as follows. Section II is devoted to an overview of error measures based on statistics. Next, Section III reviews most existing reference-based edge measures involving distances of erroneous points, then points out the advantages and drawbacks of different measures, followed by the presentation of a new measure. Section IV presents experimental results for different synthetic data, and then quality measure results for real images. Finally, Section V gives perspectives for future work and draws the conclusions of the study.

II. Confusion matrix-based error assessments

To assess an algorithm, the confusion matrix remains a cornerstone in boundary detection evaluation methods. Let G_t be the reference contour map corresponding to ground truth and D_c the detected contour map of an image I. Comparing pixel per pixel G_t and D_c, the first criterion to be assessed is the common presence of edge/non-edge points. A basic evaluation is compounded from statistics resulting from a confusion matrix. To that effect, G_t and D_c are combined. Afterwards, denoting $|\cdot|$ as the cardinality of a set, all points are divided into four sets:

- True Positive points (TPs), common points of G_t and D_c: $TP = |D_c \cap G_t|$,
- False Positive points (FPs), spurious detected edges, i.e. erroneous pixels of D_c and not in G_t defined as boundary: $FP = |D_c \cap \neg G_t|$,
- False Negative points (FNs), missing boundary points of D_c, i.e. holes in the true contour: $FN = |\neg D_c \cap G_t|$,
- True Negative points (TNs), common non-edge points of G_t and D_c: $TN = |\neg D_c \cap \neg G_t|$.

On the one hand, let us consider boundary detection of natural images, FPs appear in the presence of noise, texture or other contours influencing the filter used by the edge detection operator. On the other hand, FNs represent holes in a contour of D_c (generally caused by blurred edges in the original image I). For example, an incorrect threshold of the segmentation could generate both FPs and FNs. In the experiment, TPs, FPs and FNs are respectively represented in green, red and blue: see illustrations in Fig. 2 for more details. Computing only FPs and FNs enables a segmentation assessment to be performed [35][36]. Yet, combining at least these two quantities enables

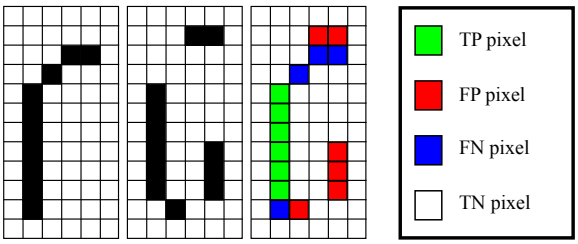


Fig. 2. Illustration of TP (green), FP (red) and FN (blue) points. In (b), D_c is contaminated with 6 FPs ans 4 FNs, illustrated with colors in (c).

TABLE I
LIST OF ERROR MEASURES INVOLVING ONLY STATISTICS.

Error measure name	Formulation	Parameters
Binary noise-to-signal ratio [58]	$B_{SNR}(G_t, D_c) = \sqrt{\frac{\lvert D_c\rvert}{FP+FN}}$	None
Complemented $Performance\ \ measure$ [51] [19]	$P_m^*(G_t, D_c) = 1 - \frac{TP}{TP+FP+FN}$	None
Segmentation Success Ratio [55]	$SSR(G_t, D_c) = 1 - \frac{TP^2}{\lvert G_t\rvert \cdot \lvert D_c\rvert}$	None
Complemented Φ measure [56]	$\Phi^*(G_t, D_c) = 1 - \frac{TPR \cdot TN}{TN+FP}$	None
Complemented χ^2 measure [60]	$\chi^{2*}(G_t, D_c) = 1 - \frac{TPR - TP - FP}{1 - TP - FP} \cdot \frac{TP+FP+FPR}{TP+FP}$	None
Complemented F_α measure [39]	$F_\alpha^*(G_t, D_c) = 1 - \frac{P_{REC} \cdot TPR}{\alpha \cdot TPR + (1-\alpha)\cdot P_{REC}}$, with $PREC = \frac{TP}{TP+FP}$	$\alpha \in]0;1]$

a segmented image to be assessed more precisely. Thus, we present a list of statistical measures -others are detailed in [3]- , and a good-quaity edge detection method should obtain the smallest response for the three following indicators [2] [9]:

$$\begin{cases} \text{Over-detection error} : Over(G_t, D_c) = \frac{FP}{\lvert I\rvert - \lvert G_t\rvert}, \\ \text{Under-detection error} : Under(G_t, D_c) = \frac{FN}{\lvert G_t\rvert}, \\ \text{Localization-error [29]:} \quad Loc(G_t, D_c) = \frac{FP+FN}{\lvert I\rvert}. \end{cases}$$

One of the pioneer works in quantitative edge evaluation, reported by Deutsch and Fram [12] computed two parameters P_1 and P_2. The first evaluates the distribution of false positive points in function of the number of columns containing edge points, whereas P_2 quantifies the edge detection by counting the missing points for each line of D_c. Thus, P_1 and P_2 parameters are given by:

$$P_1(G_t, D_c) = 1 - \frac{n_{sig}}{n_{sig} + (n_{noise} + FP) \cdot \frac{w_{stan} \cdot \lvert G_t\rvert}{w_1 \cdot \lvert I\rvert}}. \quad (1)$$

and

$$P_2(G_t, D_c) = 1 - \frac{\frac{n_r}{w_2} - \left(1 - \left[1 - \frac{n_{noise}}{\lvert G_t\rvert}\right]^{w_1}\right)}{\left[1 - \frac{n_{noise}}{\lvert G_t\rvert}\right]^{w_1}}, \quad (2)$$

with:

$$\begin{cases} n_{noise} = \frac{FP \cdot \lvert G_t\rvert}{TN+FP} \\ n_{sig} = \frac{TP - n_{noise}}{1 - n_{noise}}, \end{cases}$$

where n_r and w_1 represent respectively the number of rows and columns in D_c which contains a least one edge point (FP or TP point). Also, w_{stan} corresponds to the number of columns of G_t (which is the same as D_c), w_2 is the number of rows of the envelop rectangle containing D_c. These two parameters are close to 0 when the segmentation is efficient and tend towards one for poor detection. However, P_1 and P_2 would be more suited for the evaluation of vertical edges (see Fig. 3). On the contrary, the score of P_1 can be over 1 when $w_1 \ll w_{stan}$, as in Fig. 5. Moreover, concerning P_2, when $n_r = w_2$, so $\frac{n_r}{w_2} = 1$ and $P_2(G_t, D_c) = 1 - \frac{(1-n_{noise}/\lvert G_t\rvert)^{w_1}}{(1-n_{noise}/\lvert G_t\rvert)^{w_1}} = 0$, translating, wrongly, a perfect segmentation, as in Fig. 3 (b), (c), (e), (f) and Fig. 5.

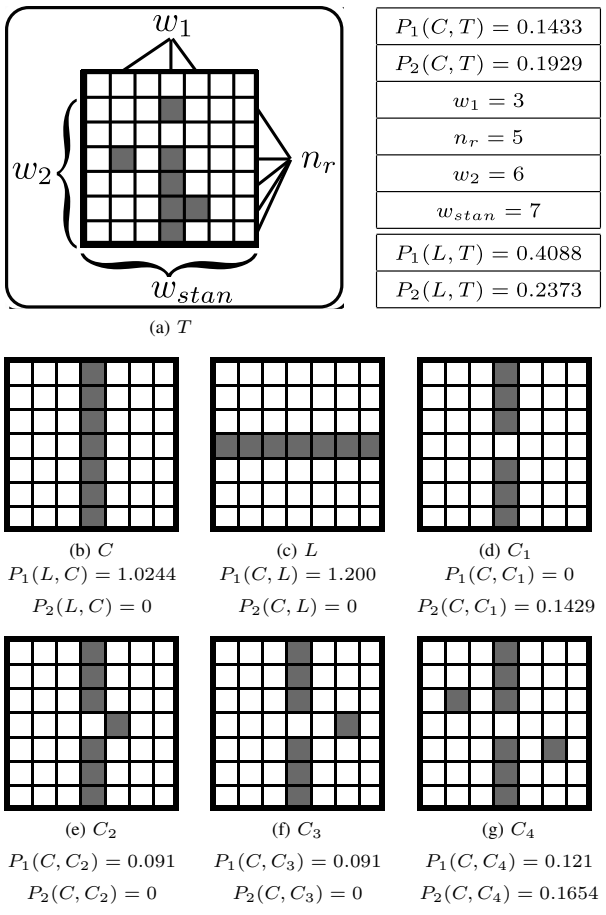


Fig. 3. Computation of parameters P_1 and P_2 using contour images 7×7. The C image in (b) is considerate as the G_t image. P_1 penalizes D_c having FPs whereas P_2 penalizes vertical contours with hole(s). In (d), $P_1(C, C_1) = 0$ because P_1 records only FPs. Note in (a) that $P_1(L, C) \neq P_1(C, L)$.

Several edge detection evaluations involving confusion matrices are presented in Table I. Only TPs are an indicator, as for SSR, this measure is normalized in function of $|G_t| \cdot |D_c|$ and decreases with improved quality of detection, with $SSR = 0$ qualifying a perfect segmentation.

The Binary Signal-to-Noise Ratio [58] B_{SNR} (see table I) is inspired by the Signal-to-Noise Ratio (SNR). It corresponds to a measure comparing the level of a signal in a reference signal to the level of the noise in a desired signal. Thus, B_{SNR} computes a global error in function of the FPs and FNs; the fewer the numbers of FPs and FNs, the more the score increases and tends to the infinity when $G_t = D_c$.

Finally, the complemented *Performance measure* P_m^* presented in table I considers directly and simultaneously the three entities TP, FP and FN to assess a binary image [19]. The measure is normalized and decreases with improved quality of detection, with $P_m^* = 0$ qualifying perfect segmentation.

By combining FP, FN, TP and TN, another way to display evaluations is to create Receiver Operating Characteristic (ROC) [7] curves or Precision-Recall (PR) [39], involving *True Positive Rates* (TPR) and *False Positive Rates* (FPR):

$$\begin{cases} \text{TPR} = \dfrac{TP}{TP+FN} \\ \text{FPR} = \dfrac{FP}{FP+TN}. \end{cases}$$

Derived from TPR and FPR, the three measures Φ, χ^2 and F_α (detailed in Table I) are frequently used in edge detection assessment. Using the complement of these measures results in a score close to 1 indicating good segmentation, and a score close to 0 indicating poor segmentation. Among these three measures, F_α remains the most stable because it does not consider the TNs, which are dominant in edge maps. Indeed, taking into consideration TN in Φ and χ^2 influences solely the measurement (as is the case in huge images).

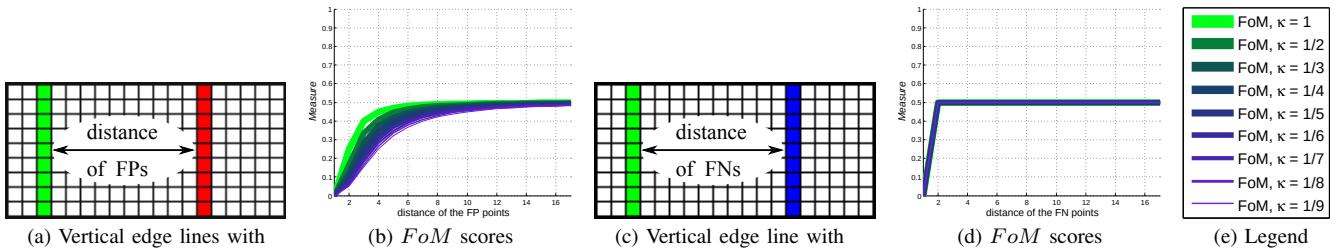
These measures evaluate the comparison of two edge images, pixel per pixel, tending to severely penalize an (even slightly) misplaced contour. Furthermore, they depend heavily of the reference image G_t which could be interpreted in different ways, as the different contours presented in Fig. 1. So statistical measures do not indicate enough significant variations of the desired contour shapes in the course of an evaluation (as illustrated in Fig. 5). As this penalization tends to be too severe, some evaluations resulting from the confusion matrix recommend incorporating spatial tolerance, particularly for the assimilation of TPs [48] [7] [39]. This inclusion could be carried by a distance threshold or a dilation of D_c and/or G_t, as in [10] (see Eq. (58)). Such a strategy of assimilation leads to counting several near contours as stripes parallel to the desired boundary (issued from the edge detector itself or a blur/texture in the original image). For example, awarding a spatial tolerance for the FPs will ensure the same score for an over-segmentation near the edges (experiments illustrated in Fig. 13), if the spatial tolerance stays greater than the FP distances from the true pixel. Tolerating a distance from the true contour and integrating several TPs for one detected contour are opposite to the principle of unicity in edge detection expressed by the 3rd Canny criterion: an optimal edge detector must produce a single response for one contour [8]. Thus, from the discussion below, only one FP or one TP should be considered in the boundary detection evaluation process. Finally, to perform an edge evaluation, the assessment should penalize a misplaced edge point proportionally to the distance from its true location.

III. Assessment involving distances of misplaced pixels

A reference-based edge map quality measure requires that a displaced edge should be penalized in function not only of FPs and/or FNs but also of the distance from the position where it should be located. Table II reviews the most relevant measures in the literature with a new one called Ψ. Some distance measures are specified in the evaluation of over-segmentation (i.e. presence of FPs) and others in the assessment of under segmentation (i.e. missing ground truth points). A complete edge detection evaluation measure takes into account both under- and over-segmentation assessment, the studied measures are detailed in the following subsection.

A. Existing quality measures involving distances

The common feature between these evaluators corresponds to the error distance $d_{G_t}(p)$ or/and $d_{D_c}(p)$. Indeed, for a pixel belonging to the desired contour $p \in D_c$, $d_{G_t}(p)$ represents the minimal distance between p and G_t. On the contrary, if a pixel p belongs to the ground truth G_t, $d_{D_c}(p)$ is the minimal distance between p and D_c. Mathematically, denoting (x_p, y_p) and (x_t, y_t) the pixel coordinates of two points p and t respectively, thus $d_{G_t}(p)$ and $d_{D_c}(p)$ are described by:



(a) Vertical edge lines with false positive points (b) FoM scores (c) Vertical edge line with false negative points (d) FoM scores (e) Legend

Fig. 4. Evolution of FoM in function of the the distance of the false positive/negative points and κ parameter. A vertical line of false positive points (a) or false negative points (c) is shifted by a maximum distance of 16 pixels and the measured scores are plotted in function of the displacement of this distance.

TABLE II
LIST OF ERROR MEASURES COMPARED IN THIS WORK. IN THE LITERATURE, THE MOST COMMON VALUES ARE $k = 1$ OR $k = 2$.

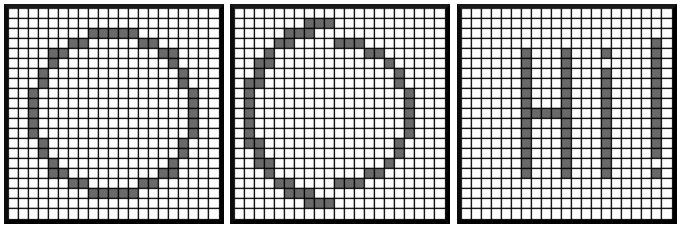
Error measure name	Formulation	Parameters
Figure of Merit (FoM) [1]	$FoM\left(G_t, D_c\right) = 1 - \frac{1}{\max\left(\left\|G_t\right|, \left|D_c\right|\right)} \cdot \sum_{p \in D_c} \frac{1}{1 + \kappa \cdot d_{G_t}^2(p)}$	$\kappa \in]0;1]$
FoM of over-segmentation [52]	$FoM_e\left(G_t, D_c\right) = 1 - \frac{1}{\max\left(\mathrm{e}^{-FP}, FP\right)} \cdot \sum_{p \in D_c \cap \neg G_t} \frac{1}{1 + \kappa \cdot d_{G_t}^2(p)}$	$\kappa \in]0;1]$
FoM revisited [47]	$F\left(G_t, D_c\right) = 1 - \frac{1}{\left\|G_t\right\| + \beta \cdot FP} \cdot \sum_{p \in G_t} \frac{1}{1 + \kappa \cdot d_{D_c}^2(p)}$	$\kappa \in]0;1]$ and $\beta \in \mathbb{R}^+$
Combination of FoM and statistics [5]	$d_4\left(G_t, D_c\right) = \frac{1}{2} \cdot \sqrt{\frac{\left(TP - \max\left(\left\|G_t\right|, \left|D_c\right|\right)\right)^2 + FN^2 + FP^2}{\left(\max\left(\left\|G_t\right|, \left|D_c\right|\right)\right)^2}} + FoM\left(G_t, D_c\right)$	$\kappa \in]0;1]$ and $\beta \in \mathbb{R}^+$
Edge map quality measure [43]	$D_p\left(G_t, D_c\right) = \frac{1/2}{\left\|I\right| - \left|G_t\right|} \cdot \sum_{p \in D_c} \left(1 - \frac{1}{1 + \kappa \cdot d_{G_t}^2(p)}\right) + \frac{1/2}{\left|G_t\right|} \cdot \sum_{p \in G_t} \left(1 - \frac{1}{1 + \kappa \cdot d_{G_t \cap D_c}^2(p)}\right)$	$\kappa \in]0;1]$
Yasnoff measure [59]	$\Upsilon\left(G_t, D_c\right) = \frac{100}{\left\|I\right|} \cdot \sqrt{\sum_{p \in D_c} d_{G_t}^2(p)}$	None
Hausdorff distance [24]	$H\left(G_t, D_c\right) = \max\left(\max_{p \in D_c} d_{G_t}(p), \max_{p \in G_t} d_{D_c}(p)\right)$	None
Distance to G_t [46] [24][13][31]	$D^k\left(G_t, D_c\right) = \frac{1}{\left\|D_c\right|} \cdot \sqrt[k]{\sum_{p \in D_c} d_{G_t}^k(p)}, \quad k = 1$ for [46]	$k \in \mathbb{R}^+$
Maximum distance [13]	$f_2 d_6\left(G_t, D_c\right) = \max\left(\frac{1}{\left\|D_c\right|} \cdot \sum_{p \in D_c} d_{G_t}(p), \frac{1}{\left|G_t\right|} \cdot \sum_{p \in G_t} d_{D_c}(p)\right)$	None
Oversegmentation [20][40]	$\Theta\left(G_t, D_c\right) = \frac{1}{FP} \cdot \sum_{p \in D_c} \left(\frac{d_{G_t}(p)}{\delta_{TH}}\right)^k, \quad k = \delta_{TH} = 1$ for [20]	for [40]: $k \in \mathbb{R}^+$ and $\delta_{TH} \in {\mathbb{R}_*}^+$
Undersegmentation [20][40]	$\Omega\left(G_t, D_c\right) = \frac{1}{FN} \cdot \sum_{p \in G_t} \left(\frac{d_{D_c}(p)}{\delta_{TH}}\right)^k, \quad k = \delta_{TH} = 1$ for [20]	for [40]: $k \in \mathbb{R}^+$ and $\delta_{TH} \in {\mathbb{R}_*}^+$
Baddeley's Delta Metric [2]	$\Delta^k(G_t, D_c) = \sqrt[k]{\frac{1}{\left\|I\right|} \cdot \sum_{p \in I} \left\|w(d_{G_t}(p)) - w(d_{D_c}(p))\right\|^k}$	$k \in \mathbb{R}^+$ and a convex function $w : \mathbb{R} \mapsto \mathbb{R}$
Symmetric distance [13][31]	$S^k\left(G_t, D_c\right) = \sqrt[k]{\frac{\sum_{p \in D_c} d_{G_t}^k(p)) + \sum_{p \in G_t} d_{D_c}^k(p)}{\left\|D_c \cup G_t\right\|}}, \quad k = 1$ for [13]	$k \in \mathbb{R}^+$
Magnier *et al.* measure [37]	$\Gamma(G_t, D_c) = \frac{FP+FN}{\left\|G_t\right\|^2} \cdot \sqrt{\sum_{p \in D_c} d_{G_t}^2(p)}$	None
Symmetric distance measure	$\Psi(G_t, D_c) = \frac{FP+FN}{\left\|G_t\right\|^2} \cdot \sqrt{\sum_{p \in G_t} d_{D_c}^2(p) + \sum_{p \in D_c} d_{G_t}^2(p)}$	None

$$\begin{cases} \text{for } p \in D_c : \\ d_{G_t}(p) = \text{Inf}\left\{\sqrt{(x_p - x_t)^2 + (y_p - y_t)^2}, t \in G_t\right\}, \\ \text{for } p \in G_t : \\ d_{D_c}(p) = \text{Inf}\left\{\sqrt{(x_p - x_t)^2 + (y_p - y_t)^2}, t \in D_c\right\}. \end{cases}$$

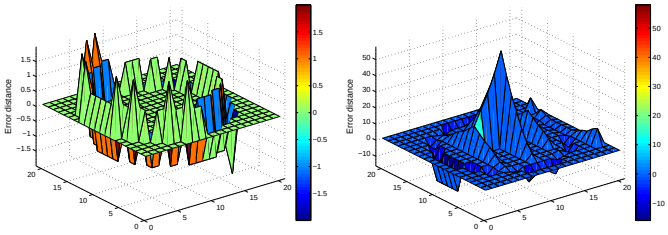
These distances refer to the Euclidean distance, even though some authors include other types of distance, see [31]. The measures presented in Table II are tuned in function of the various parameters referenced in Table IV. After marking a theoretical comparison of all these edge detection assessments and carrying out in depth comparison of their parameters, the experimental results are presented.

1) Figure of Merit and its derivatives: First, to achieve a quantitative index of edge detector performance, one of the most popular descriptors is the Figure of Merit (FoM). This distance measure ranges from 0 to 1, where 0 corresponds to a perfect segmentation [1]. The constant value κ is fixed at 1/9, 1/4 or 1/10 (the last value is the one used in the experiments of this paper). The more κ is close to 1, the more FoM tackles FPs, as illustrated in Fig 4. Widely utilized for comparing several different segmentation methods, in particular thanks to its normalization criterion, this assessment approach nonetheless suffers from a main drawback. Whenever FNs are created, for example a contour chain (even long) which is not totally extracted, the distance of FNs ($d_{D_c}(p)$) are not recorded. Indeed, FoM can be rewritten as:

$$FoM(G_t, D_c) = 1 - \frac{\sum_{p \in D_c \cap G_t} \frac{1}{1 + \kappa \cdot d_{G_t}^2(p)} + \sum_{p \in D_c \cap \neg G_t} \frac{1}{1 + \kappa \cdot d_{G_t}^2(p)}}{\max\left(\left\|G_t\right|, \left|D_c\right|\right)},$$



(a) G_t, $|G_t| = 48$ (b) D_1, $FP = 54$ $|D_c| = 54$ (c) D_2, $FP = 54$ $|D_c| = 54$



(d) Error distance: G_t v.s. D_1 (e) Error distance: G_t v.s. D_2

Negative values (in blue) represent d_{D_c} while positive values correspond to d_{G_t}.

$Over(G_t, D_{\{1,2\}}) = 0.12$	$Under(G_t, D_{\{1,2\}}) = 1$
$Loc(G_t, D_{\{1,2\}}) = 0.24$	$B_{SNR}(G_t, D_{\{1,2\}}) = 0.73$
$P_m^*(G_t, D_{\{1,2\}}) = 1$	$\Phi^*(G_t, D_{\{1,2\}}) = 1$
$\chi^{2*}(G_t, D_{\{1,2\}}) = 0.98$	$F^*_{\alpha=0.5}(G_t, D_{\{1,2\}}) = 1$
$P_1(G_t, D_1) = 16.25$	$P_1(G_t, D_2) = 1.63$
$P_2(G_t, D_1) = 0$	$P_2(G_t, D_2) = 0$
$FoM(G_t, D_1) = 0.22$	$FoM(G_t, D_2) = 0.60$
$FoM_e(G_t, D_1) = 0.22$	$FoM_e(G_t, D_2) = 0.60$
$F(G_t, D_1) = 0.63$	$F(G_t, D_2) = 0.75$
$d_4(G_t, D_1) = 0.84$	$d_4(G_t, D_2) = 0.89$
$D_p(G_t, D_1) = 0.51$	$D_p(G_t, D_2) = 0.54$
$SFoM(G_t, D_1) = 0.25$	$SFoM(G_t, D_2) = 0.40$
$MFoM(G_t, D_1) = 0.29$	$MFoM(G_t, D_2) = 0.40$
$\Upsilon(G_t, D_1) = 12.99$	$\Upsilon(G_t, D_2) = 37.34$
$H(G_t, D_1) = 1.41$	$H(G_t, D_2) = 7.67$
$H_{5\%}(G_t, D_1) = 1.41$	$H_{5\%}(G_t, D_2) = 6.71$
$D^k_{k=1}(G_t, D_1) = 1.06$	$D^k_{k=1}((G_t, D_2) = 3.05$
$D^k_{k=2}(G_t, D_1) = 0.16$	$D^k_{k=2}((G_t, D_2) = 0.48$
$f_2 d_6(G_t, D_1) = 1.06$	$f_2 d_6(G_t, D_2) = 3.05$
$\Theta_{\delta_{TH}=1}(G_t, D_1) = 1.19$	$\Theta_{\delta_{TH}=1}(G_t, D_2) = 12.26$
$\Theta_{\delta_{TH}=5}(G_t, D_1) = 0.05$	$\Theta_{\delta_{TH}=5}(G_t, D_2) = 0.49$
$\Omega_{\delta_{TH}=1}(G_t, D_1) = 1.04$	$\Omega_{\delta_{TH}=1}(G_t, D_2) = 5.12$
$\Omega_{\delta_{TH}=5}(G_t, D_1) = 0.04$	$\Omega_{\delta_{TH}=5}(G_t, D_2) = 0.20$
$\Delta^k_w(G_t, D_1) = 0.96$	$\Delta^k_w(G_t, D_2) = 2.31$
$SD^k_{k=1}(G_t, D_1) = 1.04$	$SD^k_{k=1}((G_t, D_2) = 2.57$
$SD^k_{k=1}(G_t, D_1) = 1.05$	$SD^k_{k=2}((G_t, D_2) = 2.98$
$\Gamma(G_t, D_1) = 0.34$	$\Gamma(G_t, D_2) = 0.57$
$\Psi(G_t, D_1) = 0.46$	$\Psi(G_t, D_2) = 0.72$
$KPI_\Gamma(G_t, D_1) = 0.16$	$KPI_\Gamma(G_t, D_2) = 0.27$
$KPI_\Psi(G_t, D_1) = 0.22$	$KPI_\Psi(G_t, D_2) = 0.37$

Fig. 5. Results of evaluation measures. For the two candidate images, the number of FPs and number of FNs are the same: FPs: $|D_1 \cap \neg G_t| = |D_2 \cap \neg G_t| = 54$ and FNs: $|\neg D_1 \cap G_t| = |\neg D_2 \cap G_t| = |G_t| = 48$. Also, $D_1 \cap G_t = D_2 \cap G_t = \varnothing$, so $TP = 0$ and $SSR(G_t, D_1) = SSR(G_t, D_2) = 1$. The assessments involving distances of FPs and/or FNs heavily penalize D_c, including misplaced points with greater distances.

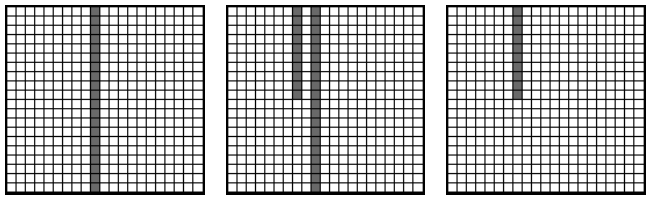
hence:

$$FoM(G_t, D_c) = 1 - \frac{TP + \sum_{p \in D_c \cap \neg G_t} \frac{1}{1 + \kappa \cdot d^2_{G_t}(p)}}{\max(|G_t|, |D_c|)}, \quad (3)$$

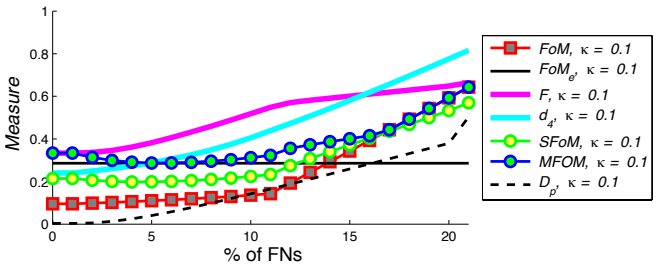
because, for $p \in D_c \cap G_t$, $d^2_{G_t}(p) = 0$ and $\frac{1}{1+\kappa \cdot d^2_{G_t}(p)} = 1$. Knowing that $TP = |G_t| - FN$, for the extreme cases, the FoM measures takes the following values:

$$\begin{cases} \text{if } FP = 0\text{:} \\ FoM(G_t, D_c) = 1 - \dfrac{TP}{|G_t|}, \\ \text{if } FN = 0\text{:} \\ FoM(G_t, D_c) = 1 - \frac{1}{\max(|G_t|, |D_c|)} \cdot \displaystyle\sum_{p \in D_c \cap \neg G_t} \frac{1}{1 + \kappa \cdot d^2_{G_t}(p)}. \end{cases}$$

Consequently, FoM counts only TPs to penalize FN points (Fig 4 (d)) whereas only distances of FPs are recorded (Fig 4 (b)). Moreover, for $FP > 0$, as $\frac{1}{1+\kappa \cdot d^2_{G_t}(p)} < 1$, it can be easily demonstrate that the FoM measure penalizes the over-detection very low compared to the under-detection. The curve in Fig. 6 shows that when $FN > FP$, the penalization of missing points (FNs) becomes higher whereas it is weak $FN < FP$. Incidentally, FoM_e represents an extension of FoM by counting only FPs [52], strictly evaluating the over-segmentation. In fact, it computes a mean of $\frac{1}{1+\kappa \cdot d^2_{G_t}(p)}$ for all the FPs. Consequently, in the presence or absence of FNs, a contour image is considered by the FoM_e criterion as correctly segmented. The experiments in Fig. 6, Fig. 8 and Section IV illustrate this undesirable behavior. Some improvements have been developed, such as F and d_4. In order to overcome the gaps in FoM, F computes the distances of FNs from G_t and the measure is tuned by the number of FPs and the cardinality of G_t (choosing $\beta = 1$ [47]). As



(a) Image G_t, $|G_t| = 21$ (b) Image L_1, $|D_c| = 31$, $TP =, 21$ $FP = 10$, $FN = 0$ (c) Image L_2, $|D_c| = 10$, $TP = 0$, $FP = 10$, $FN = 21$



(d) Scores of FoM and its derived measures.

Fig. 6. Results of FoM and its derived measures in function of the % of FNs, comparing G_t and L_1. When the number of FNs attains 100%, the candidate edge image corresponds to L_2.

TABLE III
BEHAVIOR OF d_4 MEASURE IN FUNCTION OF DIFFERENT SEGMENTATION LEVELS, NOTING $M = (\max(|G_t|, |D_c|))^2$.

Segmentation level	FoM	$\frac{\lvert TP - \max(\lvert G_t\rvert, \lvert D_c\rvert)\rvert}{M}$	$\frac{FN}{M}$	$\frac{FP}{M}$
Good	≈ 0	≈ 0	≈ 0	≈ 0
$FP \nearrow$, $FN = 0$	> 0	> 0	0	> 0
$FP = 0$, $FN \nearrow$	> 0	> 0	> 0	0
$FP \nearrow$, $FN \nearrow$	> 0	> 0	> 0	> 0

pointed out through Eq. 3 for FoM and illustrated in Fig. 6, F behaves inversely to FoM concerning FPs/FNs points: F is more sensitive to FPs than FNs. Also, d_4 represents another enhancement, this edge measure depends particularly on TP, FP, FN and FoM; d_4 is normalized with the $\frac{1}{2}$ coefficient [5]. Nonetheless, even though d_4 behaves correctly for the experiment in Fig. 6, this measure focuses on FPs and penalizes FNs like the FoM measure, as detailed in Table III. This measure suffers from two main drawbacks. On the one hand, the sum of all its terms yields a high sensibility to edge displacements. On the other hand, in the case of a pure under-segmentation, when $FN \to 0$, d_4 belongs to [0.7, 0.85] but does not attains the score of 1.

As described here, an effective measure for the evaluation of the edge detection is not only directed by d_{G_t} or d_{D_c}. Thus, inspired by $f_2 d_6$ (see bottom and [13]), another way to avoid the computation of only the distance of FPs in FoM (or only the distance of FNs in F) is to consider the combination of both $FoM\,(G_t, D_c)$ and $FoM\,(D_c, G_t)$, as in the following two formulas:

- Symmetric Figure of Merit:

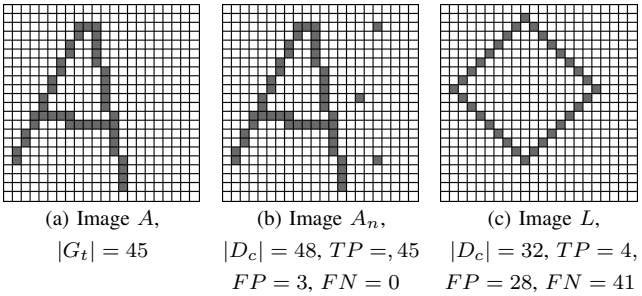
$$SFoM\,(G_t, D_c) = \frac{1}{2} \cdot FoM\,(G_t, D_c) + \frac{1}{2} \cdot FoM\,(D_c, G_t) \tag{4}$$

- Maximum Figure of Merit:

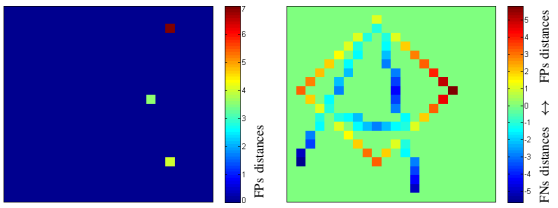
$$MFoM\,(G_t, D_c) = \max\left(FoM\,(G_t, D_c), FoM\,(D_c, G_t)\right). \tag{5}$$

As FoM is a normalized measure, $SFoM$ and $MFoM$ are also normalized. Finally, $SFoM$ and $MFoM$ take into account both distances of FNS (i.e. d_{D_c}) and FPs (i.e. d_{G_t}), so they can compute a global evaluation of a contour image. Nevertheless, Fig. 6 shows that $SFoM$ behaves like FoM when $FN > FP$ and that $MFoM$ is not monotonous for this experiment (whereas only FNs are added).

2) Hausdorff distance its enhancements: A second measure widely computed in matching techniques is represented by the Hausdorff distance H. In object recognition [22], the algorithm aims to minimize H, which measure the mismatch of two sets of points [24][45]. This max-min distance could be strongly deviated by only one pixel which can be positioned sufficiently far from the pattern (illustrated in Fig. 8); so the measured distance becomes that between the pattern and the (erroneous) point, in that case disturbing the score of H. To improve the measure such that H becomes less sensitive to outliers, one idea is to compute H with a proportion of the maximum distances (for example 5%, 10% or 15% of the values [24]); let us note $H_{n\%}$ this measure for $n\%$ of values



(a) Image A, $|G_t| = 45$ (b) Image A_n, $|D_c| = 48$, $TP =, 45$ $FP = 3$, $FN = 0$ (c) Image L, $|D_c| = 32$, $TP = 4$, $FP = 28$, $FN = 41$



(d) Distance map: A vs. A_n (e) Distance map: A vs. L

$P_1(A, A_n) = 0.01$	$P_1(A, L) = 0.80$
$P_2(A, A_n) = 0$	$P_2(A, L) = 0$
$P^*_m(A, A_n) = 0.06$	$P^*_m(A, L) = 0.95$
$SSR(A, A_n) = 0.06$	$SSR(A, L) = 0.99$
$\Phi^*(A, A_n) = 0.01$	$\Phi^*(A, L) = 0.92$
$\chi^{2*}(A, A_n) = 0.07$	$\chi^{2*}(A, L) = 0.99$
$F^*_{\alpha=0.5}(A, A_n) = 0.03$	$F^*_{\alpha=0.5}(A, L) = 0.90$
$FoM(A, A_n) = 0.05$	$FoM(A, L) = 0.60$
$FoM_e(A, A_n) = 0.50$	$FoM_e(A, L) = 0.83$
$F(A, A_n) = 0.06$	$F(A, L) = 0.66$
$d_4(A, A_n) = 0.05$	$d_4(A, L) = 0.78$
$SFoM(A, A_n) = 0.06$	$SFoM(A, L) = 0.53$
$MFoM(A, A_n) = 0.06$	$SFoM(A, L) = 0.60$
$D_p(A, A_n) = 0.003$	$D_p(A, L) = 0.29$
$\Upsilon(A, A_n) = 2.93$	$\Upsilon(A, L) = 1.82$
$H(A, A_n) = 7$	$H(A, L) = 5.83$
$H_{5\%}(A, A_n) = 5.5$	$H_{5\%}(A, L) = 5.37$
$D^{k,k=1}(A, A_n) = 0.30$	$D^{k,k=1}((A, L) = 2.05$
$D^{k,k=2}(A, A_n) = 0.18$	$D^{k,k=2}((A, L) = 0.44$
$f_2 d_6(A, A_n) = 0.30$	$f_2 d_6(A, L) = 2.11$
$\Theta_{\delta_{TH}=1}(A, A_n) = 4.87$	$\Theta_{\delta_{TH}=1}(A, L) = 2.34$
$\Theta_{\delta_{TH}=3}(A, A_n) = 2.89$	$\Theta_{\delta_{TH}=3}(A, L) = 0.79$
$\Omega_{\delta_{TH}=1}(A, A_n) = 0$	$\Omega_{\delta_{TH}=1}(A, L) = 2.31$
$\Omega_{\delta_{TH}=3}(A, A_n) = 0$	$\Omega_{\delta_{TH}=3}(A, L) = 0.77$
$\Delta^k(A, A_n) = 2.11$	$\Delta^k(A, L) = 2.56$
$SD^{k,k=1}(A, A_n) = 2.12$	$SD^{k,k=1}(A, L) = 2.09$
$SD^{k,k=2}(A, A_n) = 2.71$	$SD^{k,k=2}(A, L) = 2.51$
$\Gamma(A, A_n) = 0.01$	$\Gamma(A, L) = 0.48$
$\Psi(A, A_n) = 0.01$	$\Psi(A, L) = 0.75$
$KPI_\Gamma(A, A_n) = 0.01$	$KPI_\Gamma(A, L) = 0.24$
$KPI_\Psi(A, A_n) = 0.01$	$KPI_\Psi(A, L) = 0.39$

Fig. 7. Results of evaluation measures. Even edge evaluations involving distances can be disturbed by few misplaced pixels and do not respect the shape of the true pattern.

($n \in {\mathbb{R}_*}^+$). Even though a mean percentage of the distance does not guarantee an optimized comparison, as illustrated in Fig. 7, this enhancement helps with regard to robustness, and some other modifications are proposed in [61] and in [4].

Inspired by the Hausdorff distance with a view to developing

a new method that is robust with regard to a small number of outliers, some researchers have proposed other measures and studied their behaviors in the presence of misplaced edge points [2] [13]. As $H_{n\%}$ remains close to the Hausdorff distance, the rank n acts as a threshold for erroneous pixels and $H_{n\%}$ behaves as H. As pointed out in [13], an average distance from the edge pixels in the candidate image to those in the ground truth is more appropriate for matching purposes than H and $H_{n\%}$. A first proposition of this distance is D^k which represents an error distance only in function of d_{G_t}. Also, the Yasnoff measure, called Υ, seems to D^k, with $k = 2$, using a different coefficient of $\frac{100}{|I|}$ [59]. The distance measures D^k and Υ estimate the divergence of FPs; in other words, they correspond to a measure of over-segmentation. On the contrary, the sole use of a distance d_{D_c} instead of d_{G_t} enables an estimation of the FN divergences, representing an under-segmentation (as in Ω). Precisely, θ and Ω represent two over- and under-segmentation assessment measures, where δ_{TH} is the maximum distance allowed to search for a contour point (see also distortion rates in [18]). These distance measures penalize misplaced points further than δ_{TH} from G_t [40]. Choosing, $\delta_{TH} > 1$ is equivalent to taking into account the relative position for the over- and under-segmentation and goes back to the problem of the spatial tolerance detailed in Section II). Finally, as concluded in [9], a complete and optimum edge detection evaluation measure should combine assessments of both over- and under-segmentation, as f_2d_6 and S^k. Thus, the score of the f_2d_6 corresponds to the maximum between the over- and the under-segmentation whereas the values obtained by S^k reprensents their mean. Moreover, S^k takes small values in the presence of low level of outliers whereas the score becomes large as the level of mistaken points increases [13][31] but is sensitive to remote misplaced points as represented in Fig. 8.

Combining both d_{D_c} and d_{G_t}, Baddeley's Delta Metric (Δ^k) [2] is a measure derived from the Hausdorff distance which is intended to estimate the dissimilarity between each element of two binary images. Finally, as this distance measure is based on the mean difference between the two compared images and is useful in region segmentation [26]. For this measure, w represents a weighting concave function, in general $w(x) = x$ or $w(x) = \min(\sqrt{n^2 + m^2}, x)$ for an image of size $m \times n$, with m and $n \in \mathbb{N}^*$ [30] (in our experiments, we use $w(x)=x$). Compared to automatic threshold algorithms such as [41], the threshold at which the edge detector obtains the best edge map is more appropriate when it is computed by the minimum value of Δ^k [14]. The main drawback of Δ^k is its hypersensitivity to false positive points, i.e. this measure tends to over-penalize images with false detections. Indeed, when a false positive pixel is far from the true edge, the $|w(d_{G_t}(p)) - w(d_{D_c}(p))|$ value creates a high impact on the evaluation, thus penalizing the measure (as in the example in Fig. 8).

Another way to compute a global measure is presented in [43] with the normalized edge map quality measure D_p. In fact, this distance measure is similar to $SFoM$, with different coefficients. The over-segmentation measure (left term) evaluates d_{D_c}, the distances between the FPs and G_t. The under-segmentation measure (right term) computes the distances of the FNs between the closest correctly detected edge pixel, i.e. $G_t \cap D_c$. That means that FNs and their distances are not counted without the presence of TP(s), and D_p is sensitive to displacements of edges. Moreover, both the left and the right terms are composed of a $\frac{1}{2}$ coefficient, so in the presence of only under- or over-segmentation, the score of D_p does not go above $\frac{1}{2}$.

3) A new edge detection measure evaluation: In [37], a normalized measure is developed after computation of the edge assessment Γ. The Γ function represents an over-segmentation measure which depends also of FN and FP. As this measure is not sufficiently efficient concerning FNs, Ψ is an alternative function which also considers d_{D_c} for false negative points. Inspired by S^k, Ψ uses different coefficients which change the behavior of the measure, as discussed in the next subsection.

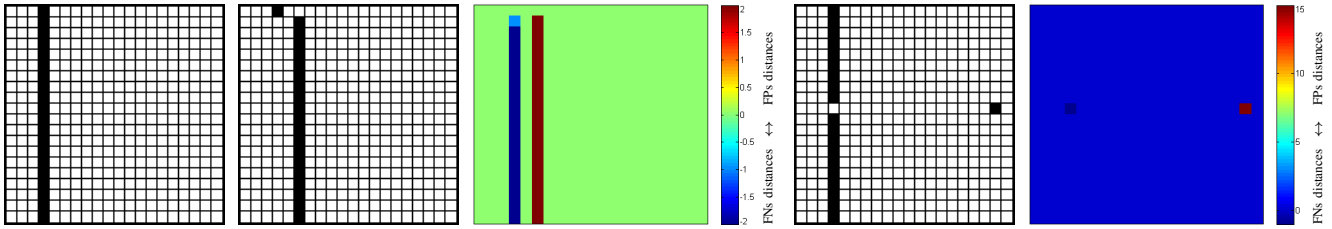
B. On the importance of the coefficients

As shown in Table IV, distance measures compute the evaluation using a coefficient mostly including: $|G_t|$, $|D_c|$, $|G_t \cup D_c|$, FN or FP. Concerning the Υ distance measure, the coefficient $\frac{100}{|I|}$ compress the measurement, especially when the image is large (as demonstrated in Fig. 12). For D_p, the coefficient $\frac{1}{|I|-|G_t|}$ affects the measure concerning the false positive term, thereby creating an insensitivity to FPs, because, generally, $|G_t| \ll |I|$and $\frac{1}{|I|-|G_t|} \sim \frac{1}{|I|} \sim 0^+$. Thus, the assessment of FNs is given more weight in the edge evaluation, strongly penalizing edge displacements. The experiments presented in the next section illustrate this drawback (Fig. 12). However, Δ^k uses $|I|$ as denominator term because the mean distance is computed for all the pixels of the image.

The authors of Γ have studied the influence of the coefficient in different concrete cases [37]. Actually, using only $|G_t|$ or $|D_c|$ penalizes severely the measurement when one misclassified pixel is placed at a significant distance of its true position.

TABLE IV
PARAMETERS AFFECTING THE ERROR DISTANCE MEASURES.

Measure	$\|G_t\|$	$\|D_c\|$	d_{G_t}	d_{D_c}	Other
FoM	✓	✓	✓		
FoM_e			✓		FP
F	✓			✓	FP
d_4	✓	✓			TP, FP, FN
$MFoM$	✓	✓	✓	✓	
$SFoM$	✓	✓	✓	✓	
D_p	✓		✓		$\|I\|$, $p \in G_t$: $d_{G_t \cap D_c}(p)$
Υ			✓		$\|I\|$
H, $H_{n\%}$			✓	✓	
D^k		✓	✓		
Θ			✓		FP
Ω				✓	FN
Δ^k			✓	✓	$\|I\|$
f_2d_6	✓	✓	✓	✓	
S^k			✓	✓	$\|G_t \cup D_c\|$
Γ	✓		✓		$FP + FN$
Ψ	✓		✓	✓	$FP + FN$



(a) Ground truth, $|G_t|$ = 20 (b) T_1 with 19 FPs, distance of 2 pixels. (c) Distance map from (a) vs. (b) (d) T_2 with 1 FP, distance of 15 pixel. (e) Distance map from (a) vs. (d)

D_c	$\lvert G_t\rvert$	$\sqrt{\sum_{p\in G_t} d^2_{D_c}(p)+\sum_{p\in D_c} d^2_{G_t}(p)}$	FN	FP	FoM_e	H	$H_{20\%}$	Δ^k	S^k	$\mathcal{A}$	$\mathcal{B}$	Ψ
T_1	100	12.21	19	19	0.5	2	2	1.91	1.93	0.031	0.031	1.15
T_2	100	15.03	1	1	0.983	15	4	4.78	2.38	0.038	0.038	0.075

Fig. 8. Errors quantified by FoM_e, H, $H_{5\%}$, Δ^k, S^k ($k=2$), $\mathcal{A}$, $\mathcal{B}$ and Γ for two different candidate edge images compared to the same ground truth.

For a counterexample, let us consider the following formula:

$$\mathcal{A}(G_t,D_c)=\frac{\sqrt{\sum_{p\in G_t} d^2_{D_c}(p)+\sum_{p\in D_c} d^2_{G_t}(p)}}{|G_t|^2}. \quad (6)$$

$\mathcal{A}$ represents an evaluation measure of contour detection involving both d_{D_c} and d_{G_t} plus a coefficient term $\frac{1}{|G_t|^2}$. Now, considering two images (images (b) and (d) in Fig. 8) with identical ground truth points G_t (Fig. 8(a)). However, in image (d), a single FP point at a distance sufficiently high behaves for $\mathcal{A}$ as several FP points on another image; these images have the same interpretation in terms of $\mathcal{A}$ measure, which is not the case with ratios 1/10 and 9/10. The table in Fig. 8 illustrates, with a simple example, this drawback for $\mathcal{A}$, Δ^k and S^k, and a similar example can be found in in Fig. 7.

Close to $\mathcal{A}$, the following measure $\mathcal{B}$ depends strongly on the desired contour number $|D_c|$:

$$\mathcal{B}(G_t,D_c)=\frac{\sqrt{\sum_{p\in G_t} d^2_{D_c}(p)+\sum_{p\in D_c} d^2_{G_t}(p)}}{|D_c|^2}. \quad (7)$$

$\mathcal{B}$ suffers from the same problem as $\mathcal{A}$. Furthermore, in general, a measure depending mainly on a coefficient $\frac{1}{|D_c|}$ (as a coefficient $\frac{1}{|I|}$) evaluates a boundary image with a value close to 0 whereas D_c stays totally misplaced, especially when $|D_c|$ is huge (when $|D_c|$ is huge, it corresponds to a total saturation of the desired contour with an inconsistent number of FPs).

Finally, the authors concluded in [37] that such a formulation must take into consideration all observable and theoretically observable cases, as illustrated in Fig. 8. That means that an effective measure has to take into account all the following input parameters $|G_t|$, $|D_c|$, FN and FP, whereas the image dimensions should not be considered. Thus, the coefficient parameter $\frac{FP+FN}{|G_t|^2}$ seems a good compromise for an evaluation measure of edge maps and has been introduced into the new formula of assessment Ψ.

C. Normalization of the edge detection evaluation

In order to compare each boundary detection assessment, all the measures must be normalized, and must also indicate the same information: an error measure close to 1 means poor segmentation whereas a value close to 0 indicates good segmentation. The values of FoM, FoM_e, F, d_4, $MFoM$, $SFoM$ and D_p comply with this $[0,1]$ condition. However, concerning the other distance measures in Table II, normalization is required. Introduced in [37], a formula called Key Performance Indicator (KPI), with $KPI\in[0,1]$ gives a value close to 1 for a poor segmentation. Alternatively, a KPI value close to 0 indicates a good segmentation. The Key Performance Indicator is defined using the following equation:

$$\begin{array}{rcl} KPI_u : [0;\infty[& \mapsto & [0;1[\\ u & \to & 1-\dfrac{1}{1+u^h}. \end{array} \quad (8)$$

where the parameter u is replaced by a distance error and h a constant such that $h\in\mathbb{R}^+_*$.

A key parameter of the KPI formula is the power of the denominator term called h. It may be called a power of observation. Inasmuch as KPI depends on its value, it evolves more or less quickly around 0.5 and embodies a range of observable cases. Average values have been determined for the error distance term $\sqrt{\sum d^2_{D_c}+\sum d^2_{G_t}}$ concerning Ψ and $\sqrt{\sum d^2_{G_t}}$ for Γ. The choice of values between 1 and 2 can be easily checked. Otherwise, the more abrupt the KPI evolution, the less the transition between 0.5 and 1 is marked (i.e. the slope of the KPI curve, for example $h=1$ in Fig. 9). Moreover, fixing $h=1$, KPI stagnates far from 1 when $\sqrt{\sum d^2_{D_c}+\sum d^2_{G_t}}$ or $\sqrt{\sum d^2_{G_t}}$ becomes high. Additionally, when $h=2$, KPI starts

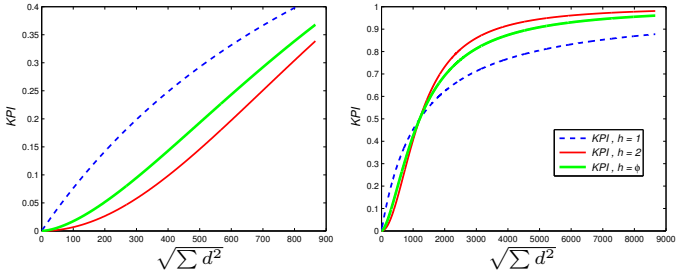
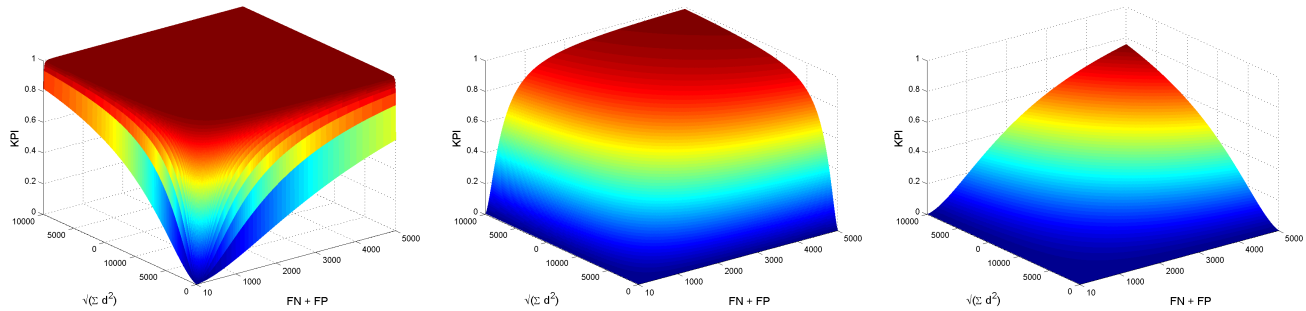


Fig. 9. Evolution of the KPI_ψ in function of the mistake points distance, for different powers h, with $FN+FP=4000$ and $|G_t|=2200$.



$|G_t| = 200$ $|G_t| = 2000$ $|G_t| = 5000$

Fig. 10. Evolution of the KPI in function of the erroneous points distance, the number of FPs and FNs, for different $|G_t|$ and fixing $h = \phi$.

to increase slowly, then the slope becomes sharp around 0.5 and converges quickly towards 1. Finally, to fix the power at the golden ratio $\phi \simeq 1.6180339887$ in order to ensure an evolution of KPI that would not be too abrupt from 0 to 1 and also not penalize when the error distances are not high (contrary to KPI with $h = 1$, see Fig. 9). Furthermore, Fig. 10 illustrates KPI values in function of the distance of the mistake points and their numbers. For a small skeleton $G_t = 200$, it shows that KPI draws near 1 quickly when the number of FPs, number of FNs and their distances are growing. Alternatively, for a bigger skeleton, KPI rises more slowly. These three evolution surfaces show the coherence and robustness of KPI regarding the starting conditions: a displaced edge is penalized in function of the number of false pixels and also of the distance from the position where it should be located.

Such a value enables a displaced edge to be penalized in function of the false pixel number and also of the distance from the position it should be located at. As a compromise, using the KPI formula with $h = \phi$, the measurement neither becomes too strong in the presence a small number of positive or negative misclassified pixels nor penalizes D_c too severely if d_{G_t/D_c} is small. Compared to Γ, Ψ improves the measurement by combining both d_{G_t} and d_{D_c}. The next section is dedicated to experimental results where candidate edge images are progressively degraded in order to study the behaviors of the presented boundary detection assessments. Note that the other measures are also normalized with another formula in the next section in order to compare all of them.

IV. Comparison of Experimental Results

The two sections above describe the main error measures concerning edge detection assessment, explaining the advantages and drawbacks of each one. The tests carried out in the experiments are intended to be as complete as possible, and thus as close as possible to reality. To that end, considering an edge model (i.e. ground truth) the edge detection evaluation measures are subjected to the following studies:

- addition of false negative points (under-segmentation),
- addition of false positive points (over-segmentation),
- addition of both false negative and false positive points,
- addition of false positive points close to the true contour,
- translation of the boundary,
- remoteness of a false positive and false negative chains,
- computation of the minimum value of the measures on edge images (synthetic and real) compared to the ground truth.

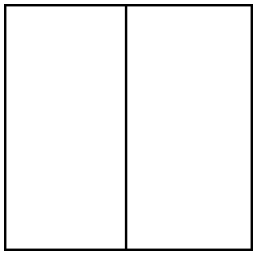
Therefore, 26 measures are tested and compared with each other: statistical measures in Table I (except B_{SNR}), distance measures in Table II, plus $SFoM$, $MFoM$ and $SSIM$. Firstly, the normalized measures Φ^*, χ^{2*}, P_m^*, F_α^*, $SSIM$, FoM, FoM_e, F, d_4, $SFoM$, $MFoM$, D_p, KPI_Γ and KPI_Ψ are plotted together. The values of SSR are not recorded because this measure behaves like P_m^*. Secondly, the scores of the other measures are also normalized to be plotted together, and normalized using the following equation for easy visual comparison. Denoting by $f \in [0; +\infty[$ the scores vector of a distance measure such that $m = \min(f)$ and $M = \max(f)$, then the normalization $\mathcal{N}$ of a measure is computed by:

$$\mathcal{N}(f) = \begin{cases} 0 & \text{if } M = m = 0 \\ 1 & \text{if } M = m \neq 0 \\ \dfrac{f - m}{M - m} & \text{if } M > 1 \text{ and } m \neq 0 \\ f & \text{otherwise.} \end{cases} \quad (9)$$

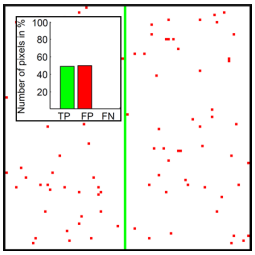
Note that the parameters for each evaluation measure are indicated directly in the captions of the curves and that the matlab code of the distance measures as FoM, D^k, S^k and Δ^k are available at http://kermitimagetoolkit.net/library/code/.

A. Behavior comparison

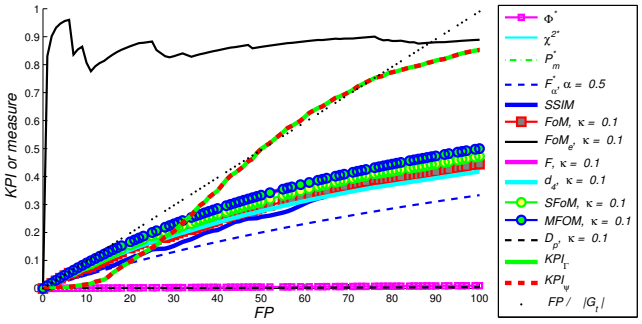
The simulation of the degradation of the ground truth G_t is studied using synthetic data. Thus, concerning the first 6 experiments, a vertical line represents G_t in a size image of 100×100, as illustrated in Fig 11(a). FNs, FPs or displacements corrupt the image, and, as G_t corresponds to a line, the interpretation of the results remains simpler. For example, adding a false pixel to G_t or a translation of G_t can easily be represented mathematically whereas, for a more complicated shape, the translation of G_t can cross other pixels and an appropriate measure will not obtain a monotonic score. Hence, the degradations applied to the vertical lines are chosen in order to obtain a monotonic score for a complete measure. Then, the following two tests concern a ground truth which is a square followed by the computation of the minimal value of the edge detection evaluation measure in



(a) G_t, image 100×100



(b) addition of 101 FPs



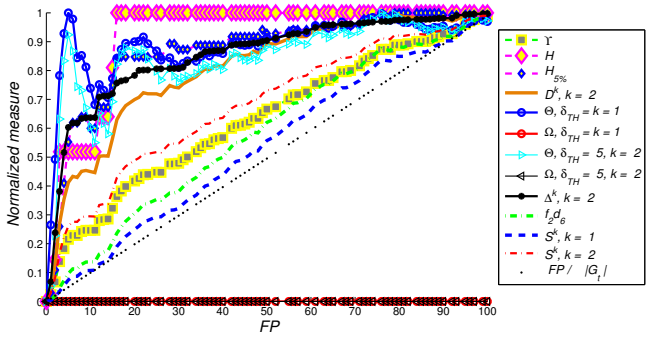
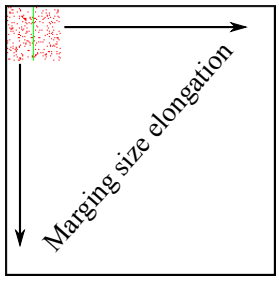
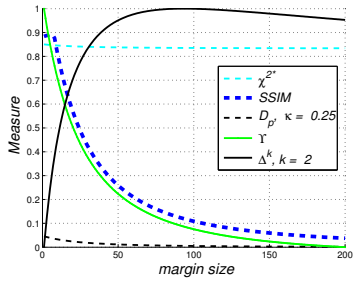


Fig. 11. Evolution of the dissimilarity measures in function of the number of added false positive points.



(a) Elongation of the margins



(b) Measure scores

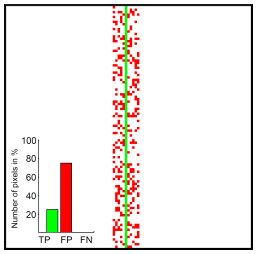
Fig. 12. Behaviors of dissimilarity measures in function of image size.

function of the threshold.

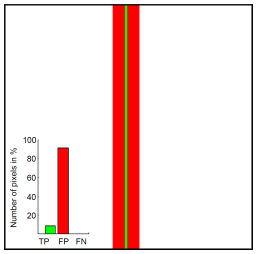
1) Addition of false positive points: The first test consisted in randomly adding undesirable pixels to G_t until 100 pixels, as represented in Fig. 11(b). This perturbation is equivalent to impulsive noise. The curves in Fig. 11 indicate the response of each measure in function of the number of created Fs; they enable comparison of the results delivered by the different measures.

Firstly, concerning measures automatically given values between 0 and 1, χ^{2*}, P_m^*, F_α^*, $SSIM$, FoM, F, d_4, $SFoM$ and $MFoM$ overlap and penalize corrupted images by FPs as soon as they appear, but their values are under 0.5 whereas the last corrupted image is composed of more FPs that TPs(see Fig. 11(b)). FoM_e is not monotonic and computes a mean in function of the apparition of FPs. Φ^* and D_p overlap and stay close to 0 (same remark for the under-segmentation Ω). Contrary to the previous measures, KPI_Γ and KPI_Ψ ensure an evaluation which does not become too high in the presence of a small number of FPs, but penalizes D_c more severely for 50 or more undesirable points. Secondly, concerning the non-normalized measures, Θ is not monotonic or sensitive to FP distances, as H and $H_{5\%}$ which stay blocked at the higher distances values, whatever the number of FPs. Furthermore, Υ, D^k, Δ^k, $f_2 d_6$ and S^k have a coherent behavior, even though Υ, D^k, Δ^k and S^k (with $k = 2$) are sensitive to FPs at the beginning of the assessment.

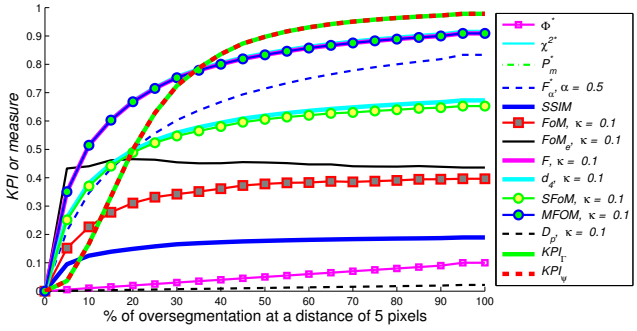
This first experiment is produced by randomly adding FPs to the same G_t image. The second test is presented in the image of Fig. 12 (a) which is composed of a vertical line and 100 FPs randomly placed around the line. The first image in this test image is as large as the image in Fig. 11(b). In a second step, the TPs and FPs are maintained in the same position and the margins are increased, as illustrated in Fig. 12 (b), influencing the evaluation in certain cases. The curves presented in Fig. 12 (b) show which dissimilarity measure



(a) 30 % of FPs



(b) 100 % of FPs



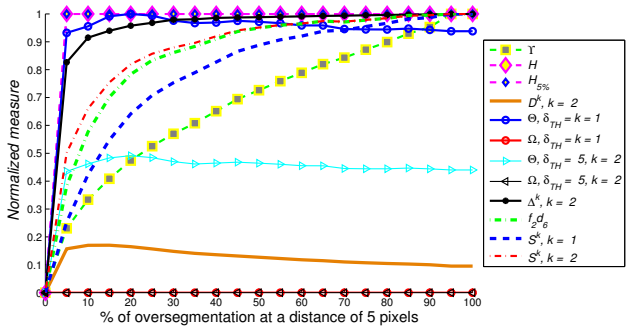


Fig. 13. Evolution of the dissimilarity measures in function of the over-segmentation in the contour area.

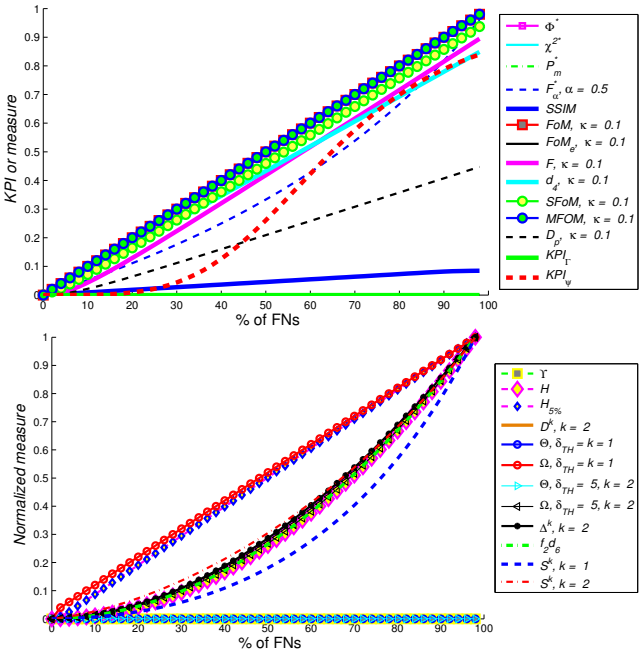


Fig. 14. Evolution of the dissimilarity measures in function of the number of false negative points addition. FoM and $MFoM$ overlap.

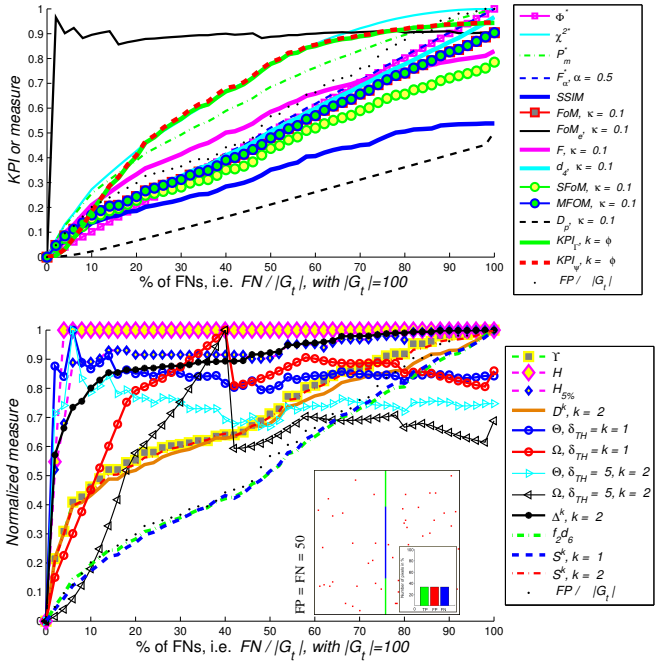


Fig. 15. Evolution of the dissimilarity measures in function of both the number of false positive and false negative points addition.

scores are changed by the elongations of the margins (note that the margins lengthen both downward and to the right at the same time). Even though the values of χ^{2*} do not change significantly, $SSIM$, Υ and Δ^k evolve strongly. The distance measure D_p, which remains insensitive to FPs becomes increasingly close to zero as the margin increases.

2) Over-segmentation close to the contour: The idea of this experiment is to create an over-segmentation at a maximum distance of 5 pixels, as illustrated in Fig. 13(a). Therefore, 100% of over-segmentation represents a dilation of the vertical line with a structural element of size 1×6, corresponding of a total saturation of the contour, see Fig. 13(b). The curves presented in Fig. 13 show that FoM_e, Υ, $S^k(k=2)$, Δ^k and Θ ($\delta_{TH}=1$) are very sensitive to FPs whereas, Φ^*, D^k, FoM_e, $SSIM$ and D_p do not penalize D_c enough. Also, Ω stagnates at 0 because it corresponds to an under-segmentation measure. FoM_e, D^k and Θ are not monotonic whereas the saturation of the contour is progressive. Note that H and $H_{5\%}$ keep the same result throughout the test after 5% of degradation and that Δ^k is nearly the same. Finally, χ^{2*}, F^*_α, FoM, F, d_4, $SFoM$, $MFoM$, KPI_Γ, KPI_Ψ, f_2d_6 and $S^k(k=1)$ ensure a measure evolution which is not too abrupt, even though FoM stagnates around 0.35, then d_4 and $SFoM$ around 0.6.

3) Addition of false negative points: In this experiment, pixels of the vertical line are missing, creating false negative points. Fig. 14 presents the evolution of the criteria obtained by the studied segmentation measures. These curves indicate that almost all the measures show the same behavior. The over-segmentation measures FoM_e, Γ, Υ, D^k and Θ stagnate at 0. The $SSIM$ stays close to 0, whereas there are almost no more remaining pixels in D_c at the end of the experiment. As this test concerns only the addition of FNs, the D_p distance measure obtains a result of 0.5 at the finish, but no higher because it corresponds to a measure which separates under- and over-segmentation. As feared in Section III, the d_4 measure does not sufficiently penalizes the under-segmentation because the maximum score it attains is around 0.8. Moreover, compared to the experiment in Fig. 11, it is clear that FoM, F, d_4, $SFoM$, $MFoM$ and D_p are more sensitive to FN addition than FP points. The

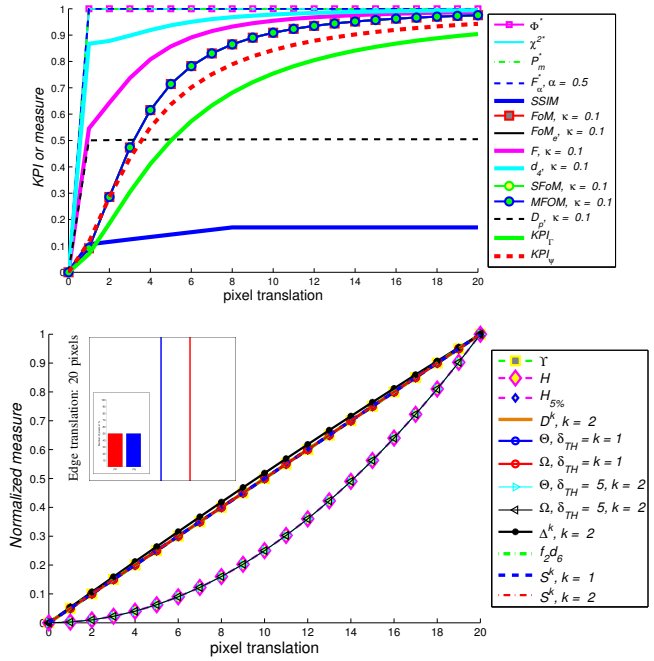
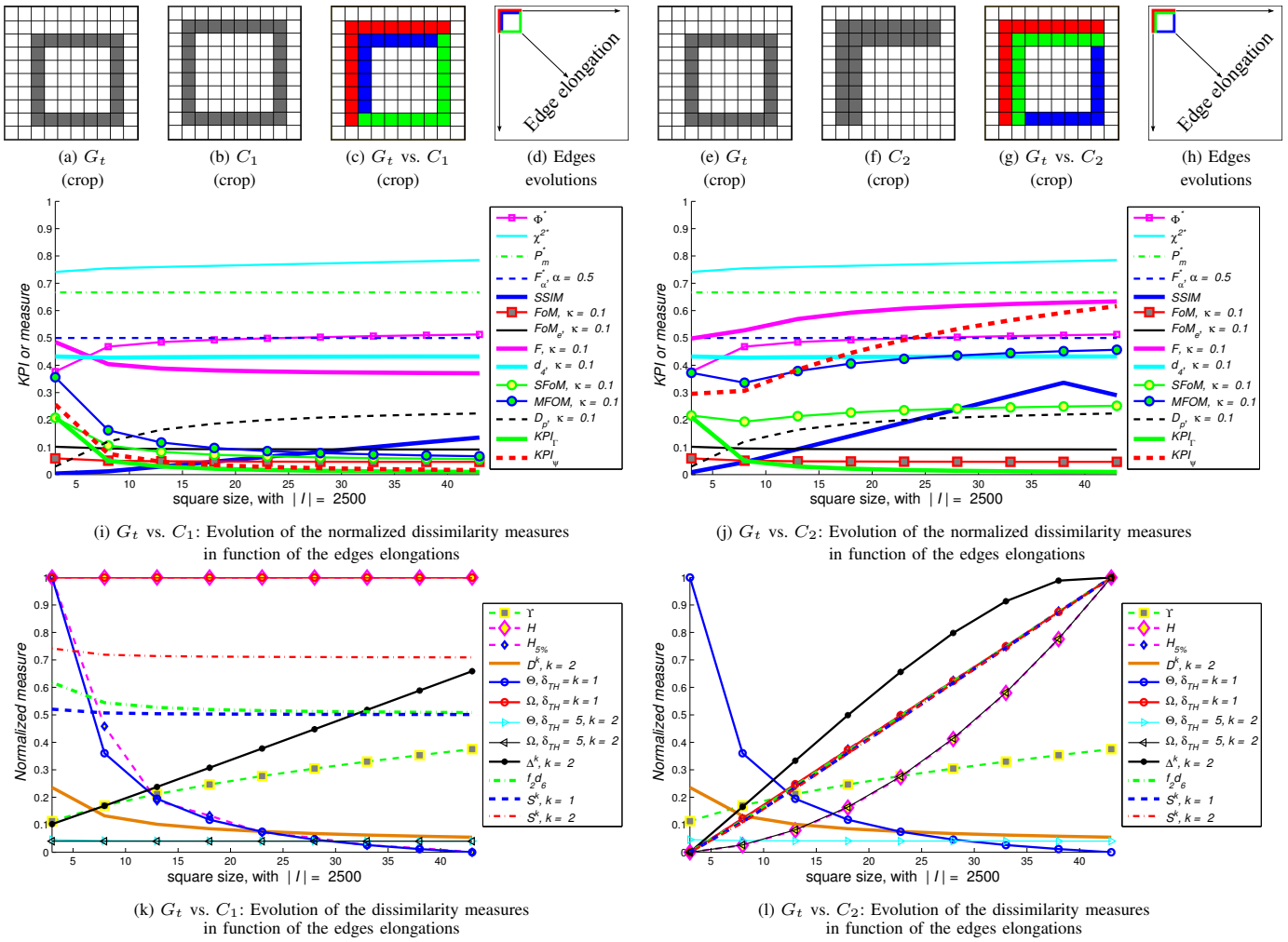


Fig. 16. Evolution of the dissimilarity measures in function of the translation of D_c. Note that FoM, FoM_e, $SFoM$ and $MFoM$ overlap.



(a) G_t (crop) (b) C_1 (crop) (c) G_t vs. C_1 (crop) (d) Edges evolutions (e) G_t (crop) (f) C_2 (crop) (g) G_t vs. C_2 (crop) (h) Edges evolutions

(i) G_t vs. C_1: Evolution of the normalized dissimilarity measures in function of the edges elongations

(j) G_t vs. C_2: Evolution of the normalized dissimilarity measures in function of the edges elongations

(k) G_t vs. C_1: Evolution of the dissimilarity measures in function of the edges elongations

(l) G_t vs. C_2: Evolution of the dissimilarity measures in function of the edges elongations

Fig. 17. Evolution of dissimilarity measures in function of the size of the original square.

non-normalized measure scores in the lower graph of Fig. 14 evolve progressively, even though S^k (with $k = 1$) is less sensitive to FNs Note that, for Δ^k (which overlaps with $f_2 d_6$), FN distances are considered to be less detrimental than FP distances (experiments reported in Fig. 11). Finally, KPI_Ψ ensures an evaluation which is not too high in the presence of a small number of FNs, followed by a greater penalization when the number of FNs increases.

4) Addition of both false positive and false negative points: Some measure are specialized and have an adaptive behavior for the evaluation of only under- or over-segmentation. Nevertheless, it is interesting to study their sensitivities by combining these two degradations because edge detectors often create both FPs and FNs in real/noisy images. Thus, adding randomly undesirable pixels to G_t (until 100 pixels) and creating FNs at the same times gives a discontinuous line in the middle of randomly placed FPs, as illustrated in Fig 15, bottom right.

Fig. 15 shows a hyper-sensitivity of FPs/FNs distances for FoM_e, H, $H_{5\%}$, Θ, Δ^k, S^k and Δ^k. Moreover, Θ and Ω are not monotonic. Also, Φ^*, F^*_α, FoM, d_4, $MFoM$, $f_2 d_6$ and S^k (with $k = 1$) increase (almost) constantly from 0 to 1 whereas both FNs and FPs are added. For example, the image in Fig 15, bottom right, illustrates the line where both 50% of the correct pixels are missing and 50 FPs are added. In this precise case, Φ^*, F^*_α, FoM, d_4, $MFoM$ and $SFoM$ measures give a result close to 0.5 whereas G_t is disturbed by more than 50%, as represented by the measurement of χ^{2*}, P^*_m, KPI_Γ and KPI_Ψ (note that Υ, D^k and S^k -with $k = 2$- behave the same way even though they have a sensitivity to misplaced points at the beginning of the experiment). F seems to evolve correctly but is not monotonic. Note that the $SSIM$ remains unsuitable for this type of degradation because the best score it attains is little more than 0.5. Finally, the evolution of D_p is monotonic until 0.5 but not thereafter, due to the coefficient $\frac{1}{|I|-|G_t|}$ which compresses the result, whereas the final image represents a cloud of FPs without TPs (the next experiment with edge translation shows more clearly the influence of this coefficient).

5) Edge translation: As pointed out in Section I, a blur in the original image can shift the detected contour. When this displacement remains not too high, the shape of the desirable object stays detectable and the evaluation should not overly penalize the edge detector. Hence, edge detection evaluation

measures must be assessed in function of the translation of a true contour. Indeed, the vertical line is shifted by a maximum distance of 20 pixels (Fig. 16 bottom left) and the score of the measures is plotted in Fig. 16 in function of the displacement of the desired contour. Thus, regarding the statistical measures Φ^*, χ^{2*}, P_m^* and F_α^*, the curves perfectly illustrate the need to consider the distances of the misplaced pixels for the edge detection evaluation. FoM, $SFoM$ and $MFoM$ overlap completely, but their evaluations are correct, like KPI_Γ and KPI_Ψ. F and d_4 measures are sensitive to the first 2 translations. As pointed out above with the previous test, D_p is not suited to the evaluation of translations, like $SSIM$ (the $SSIM$ computes locally a score in a window 8×8, so the score does not change after a switch of 8 pixels). The evolution of the other measures (Fig. 16, bottom) perform monotonically but no information can be infered because the measures are not normalized between 0 and 1 to evaluate the segmentation.

6) Comparison of the false positive distance and false negative distance points: In the experiment proposed in Fig. 17, two desired contours are compared with a ground truth, illustrating the importance of considering both the distance of the false negatives points (d_{D_c}) and the distance of the false positive points (d_{G_t}). Indeed, a square (G_t in Fig. 17 (a)) is compared with another square (Fig. 17 (b)) and also with a shape of two edges forming an angle (Fig. 17 (f)). The minimum square size for G_t is 3×3 large and the image size stays fixed at 50×50. The shapes C_1 and C_2 keep the same position as in (c) and (g) and increase in proportion to G_t. All the shapes grow at the same time, as represented in Figs. 17 (d) and (h) and the scores for each measure are plotted in Fig. 17 (i), (j), (k) and (l). Thus, the more G_t grows, the more C_1 is visually closer to G_t whereas FNs deviate strongly in the case of C_2, but keep the same percentage of FPs and FNs for each desired contour. That is the reason why statistical measures obtain the same evolutions for each shape: Φ^*, χ^{2*}, P_m^* and F_α^*. Despite these two different edge evolutions, some distance measures obtain almost the same measurements for C_1 and C_2: FoM, FoM_e, D_p, D^k, Θ (with $\delta_{TH} = 1$) and KPI_Γ. On the contrary, concerning the normalized error distance measures, $MFoM$ and KPI_Ψ increase to about 0.5 for C_2, due to around 50% of TPs whereas they converge towards 0 for C_1, since C_1 becomes visually closer to G_t. Nevertheless, $SFoM$ does not sufficiently penalize FN distances in C_2, and F is too severe with C_1. Concerning non-normalized measures, $H_{5\%}$, $f_2 d_6$ and S^k behave correctly.

B. Threshold corresponding to the minimum measure

The aim of the experiments presented here is to obtain the best edge map in a supervised way. The edges are extracted using the Canny filter [8] (i.e. isotropic Gaussian filter with a standard deviation of σ). Then, thin edges are created after the non-maximum suppression of the absolute gradient in the gradient direction η [49] (see table V) and are normalized for easier comparison. In order to study the performance of the contour detection evaluation measures, one approach is to compare each measure by varying the threshold of the thin edges computed by an edge detector[1]. Indeed, compared to a ground truth contour map, the ideal edge map for a measure corresponds to the desired contour at which the evaluation obtains the minimum score for the considered measure among the thresholded gradient images. Theoretically, this score corresponds to the threshold at which the edge detection represents the best edge map, compared to the ground truth contour map [14][9]. Since a small threshold leads to heavy over-segmentation and a strong threshold may create numerous false negative pixels, the minimum score of an edge detection evaluation should be a compromise between under- and over-segmentation.

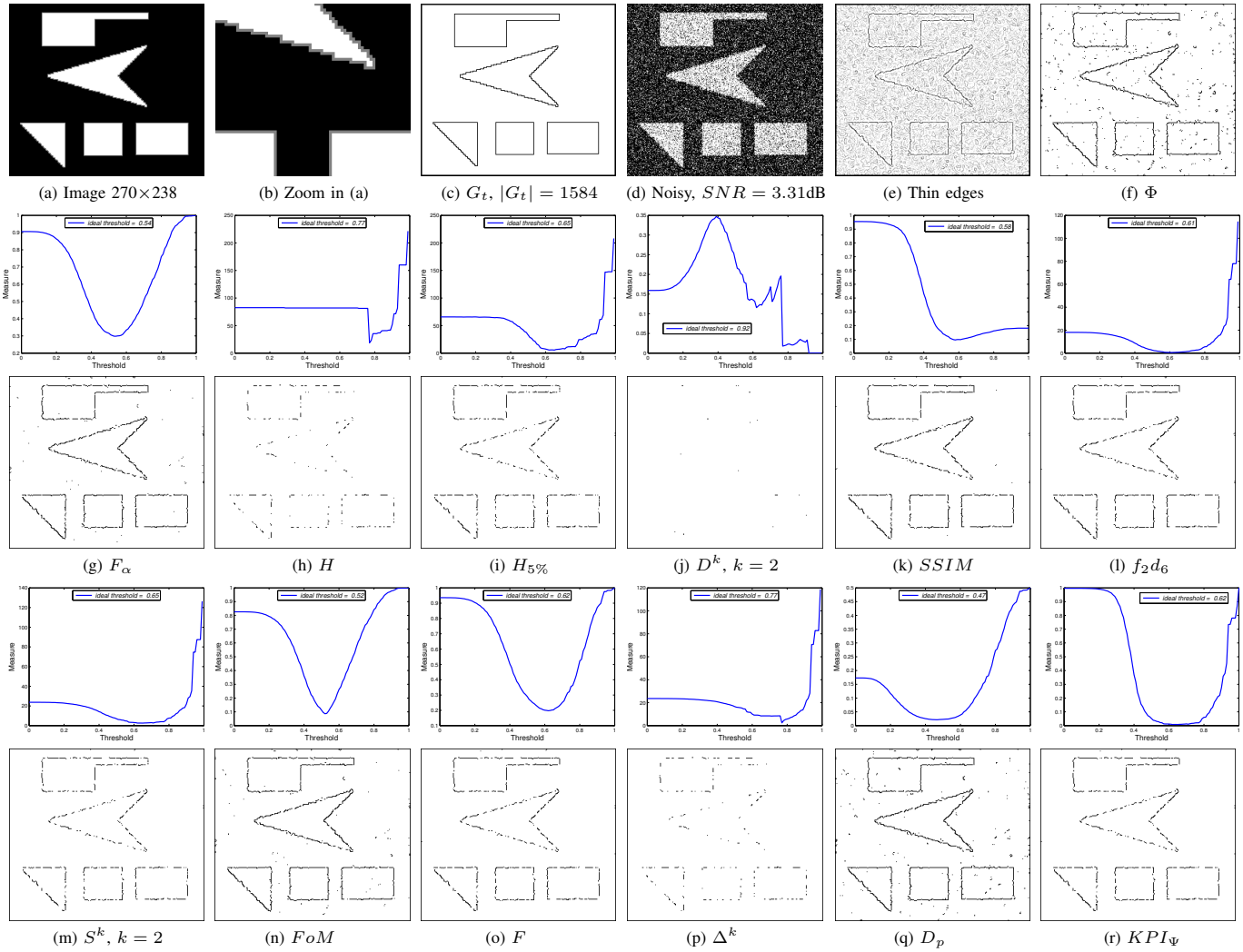
The two images used in these experiments are a synthetic and a real image. For the first one, in Fig. 18 (a), several white shapes are present immersed in a black background, creating step edges. To avoid the problem of edge pixel placements, as stated in Section I, Fig. 1, a blur is created by adding a 1 pixel width of gray around each shape. Thus, the ground truth corresponds to this gray. The second image in Fig. 19(a) is a real image with a corresponding hand-made ground truth. Even though the problem of hand-made ground truths on real images is discussed by some researchers, only the comparison of D_c with a G_t is studied here.

1) Edges of synthetic data: To evaluate the measures performances, the original image in Fig. 18(a) is disturbed with random Gaussian noise and edges are extracted from the noisy image (Fig. 18(d) and (e)). Scores are plotted in function of the threshold and the image under each curve corresponds to the ideal edge map for the considerate measure. Statistic measures and D_p correctly extract the edges at the price of numerous FPs. The Hausdorff distance H, $H_{5\%}$ and Δ^k threshold the edges too severely, losing the majority of TPs. The over-segmentation distance measures D^k, θ and Γ lose almost all the contours because the minimum corresponds to the threshold where no false pixel appears (even though true pixels disappear). Finally, $SSIM$, $f_2 d_6$, F, S^k and Ψ do not create significant holes in the contours and the FNs extracted are positioned close to the true contours.

2) Edges of a real image: Real images contain other disturbances than the previous noisy synthetic image, such as texture or blurred edges. The ground truth and the original image in Fig. 19 arise from the database available at the following website: http://www.cs.rug.nl/~imaging/databases/contour_database/contour_database.html.

In edge detection assessment, the ground truth is considered as a perfect segmentation. However, the boundary benchmarks are built by human observers and errors may be created by human labels (oversights or supplements). Indeed, an inaccurate ground truth contour map in terms of localization (sometimes several pixels, see [27], part 2.2.2.3) penalizes precise edge detectors and/or advantages the rough algorithms. In [23] the

[1]The matlab code and several measures are available on MathWorks: https://fr.mathworks.com/matlabcentral/fileexchange/63326-objective-supervised-edge-detection-evaluation-by-varying-thresholds-of-the-thin-edg



(a) Image 270×238 (b) Zoom in (a) (c) G_t, $|G_t| = 1584$ (d) Noisy, $SNR = 3.31$dB (e) Thin edges (f) Φ

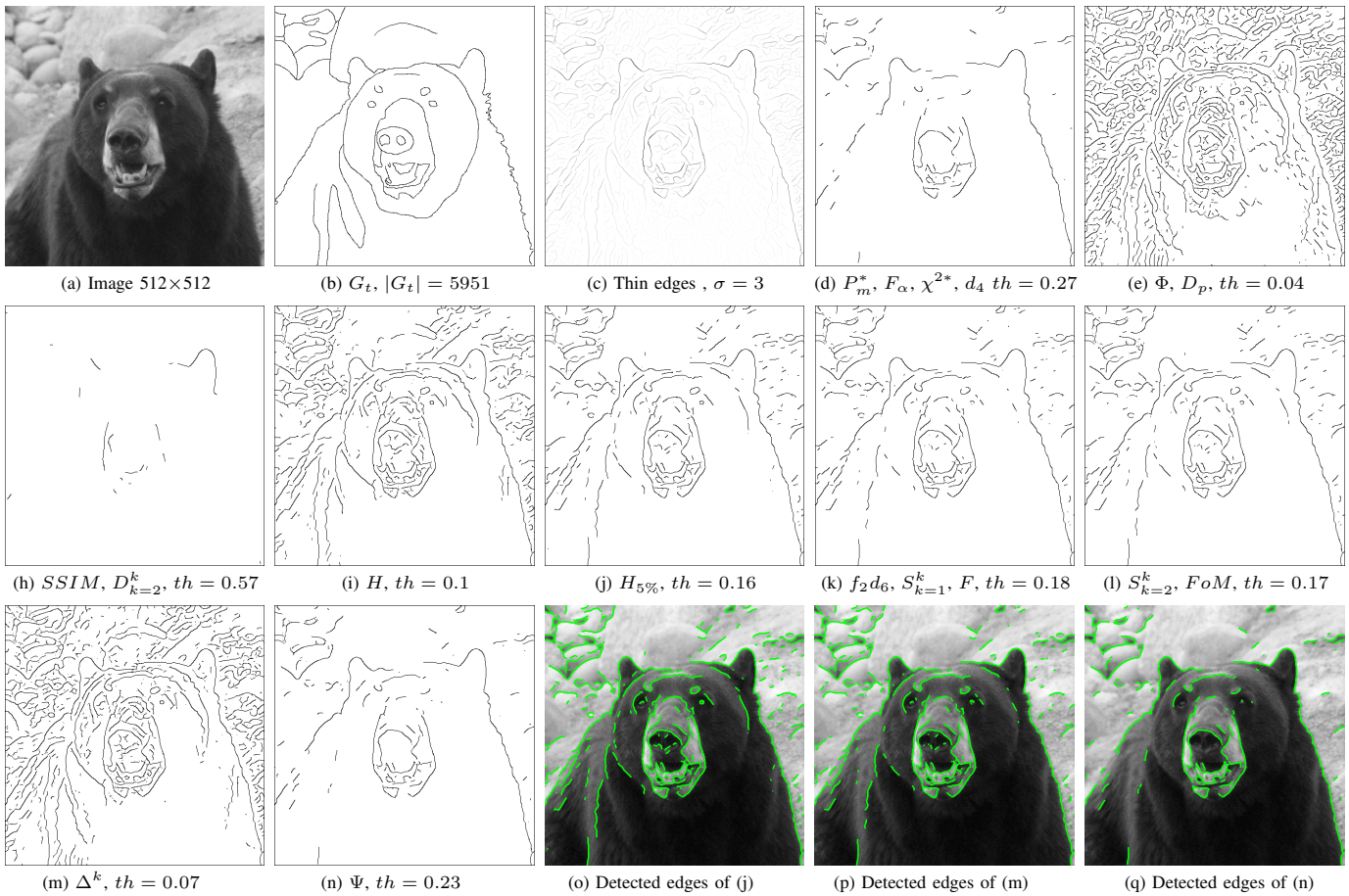
(g) F_α (h) H (i) $H_{5\%}$ (j) D^k, $k = 2$ (k) $SSIM$ (l) $f_2 d_6$

(m) S^k, $k = 2$ (n) FoM (o) F (p) Δ^k (q) D_p (r) KPI_Ψ

Fig. 18. Comparison of the minimal score for several edge detection evaluation measures concerning synthetic data. The edges are extracted using Canny's theory ($\sigma = 1$) on the noisy image in (d) and the binary images correspond to the thin edge image in (e) which is thresholded at the threshold corresponding to the minimum value of the bottom curve. The results for F_α^* are the same as the P_m^*, χ^{2*} measures and do not need to be reported. The scores of the over-segmentation measure D^k are the same as FoM_e, Υ, Θ and Γ. Moreover, FoM, d_4, $SFoM$ and $MFoM$ obtain almost the same result.

question is raised concerning the reliability of the datasets regarded as ground truths for novel edge detection methods. Thus, an incomplete ground truth penalizes a method detecting true boundaries and efficient edge detection algorithms obtain between 30% and 40% of errors. Furthermore, contrary to the previous experiment, where only noise is added to the synthetic image, the true contours are only reported by human perception and, locally, the image intensities (for example the texture) create some disturbances which are nether spatially positioned as a result of the filtering technique, nor humanly perceptible. Unfortunately, this type of contour can be amplified by a strong gradient created by the edge detector [8]. As these undesirable pixels could be far from G_t, however, an edge detection method must converge to a threshold which preserves acceptable contours, close to the ground truth, and removes undesirable contour pixels. In other words, the final shapes created by the contours of the binarized image should be similar, especially nearby the edges pixels of G_t. Thus, in this case, the segmented images are totally different from G_t than those in the synthetic case. For example, the ideal edges indicated by the $SSIM$ miss most of the main edges. Also, P_m^*, F_α^*, χ^{2*} and d_4 indicate an ideal edge image composed of numerous false positive pixels (d_4 obtains this result because most statistics include this measure). On the contrary, the statistical measure Φ, and the distance measures H, D_p and Δ^k lose most of the desired boundaries; furthermore, the ideal threshold for FoM_e and Υ is 1. Ideal contour images concerning $S_{k=2}^k$, $H_{5\%}$, $f_2 d_6$, FoM, $SFoM$ and $MFoM$ are less polluted by FPs. Finally, the ideal contour map computed for Ψ is less noisy and better-quality, leading to a compromise between optimum under- and over-segmentation.

3) Comparison of several edge detectors by filtering: In this section, we compare the score of different evaluation measures involving real images (Figs. 20(b) and 26(a)) and 9 filtering edge detectors: Sobel [53], Shen [50], Bourennane [6], Deriche [11], Canny [8], Steerable filter of order 1 (SF_1) [15], Steerable filter of order 5 (SF_5) [25], Anisotropic Gaussian Kernels (AGK) [16], Half Gaussian Kernels (H-K) [35]. Table



(a) Image 512×512 (b) G_t, $|G_t| = 5951$ (c) Thin edges , $\sigma = 3$ (d) P_m^*, F_α, χ^{2*}, d_4 $th = 0.27$ (e) Φ, D_p, $th = 0.04$

(h) $SSIM$, $D_{k=2}^k$, $th = 0.57$ (i) H, $th = 0.1$ (j) $H_{5\%}$, $th = 0.16$ (k) $f_2 d_6$, $S_{k=1}^k$, F, $th = 0.18$ (l) $S_{k=2}^k$, FoM, $th = 0.17$

(m) Δ^k, $th = 0.07$ (n) Ψ, $th = 0.23$ (o) Detected edges of (j) (p) Detected edges of (m) (q) Detected edges of (n)

Fig. 19. Comparison of the minimal score for several edge detection evaluation measures concerning a real image and a hand-made ground truth. The edges are extracted using Canny's theory ($\sigma = 3$) on the real image in (a) and the binary images correspond to the thin edge image in (c) which is thresholded at the value (th) corresponding to the minimum score for each measure. The best images for the measures $S_{k=1}^k$, $f_2 d_6$, $SFoM$ and $MFoM$ are the same. The optimal threshold for the over-segmentation measures $\Theta_{\delta_{TH}=1 \text{ or } 5}$, Γ is $th = 1$, and is $th = 0$ for the under-segmentation measure $\Omega_{\delta_{TH}=1 \text{ or } 5}$.

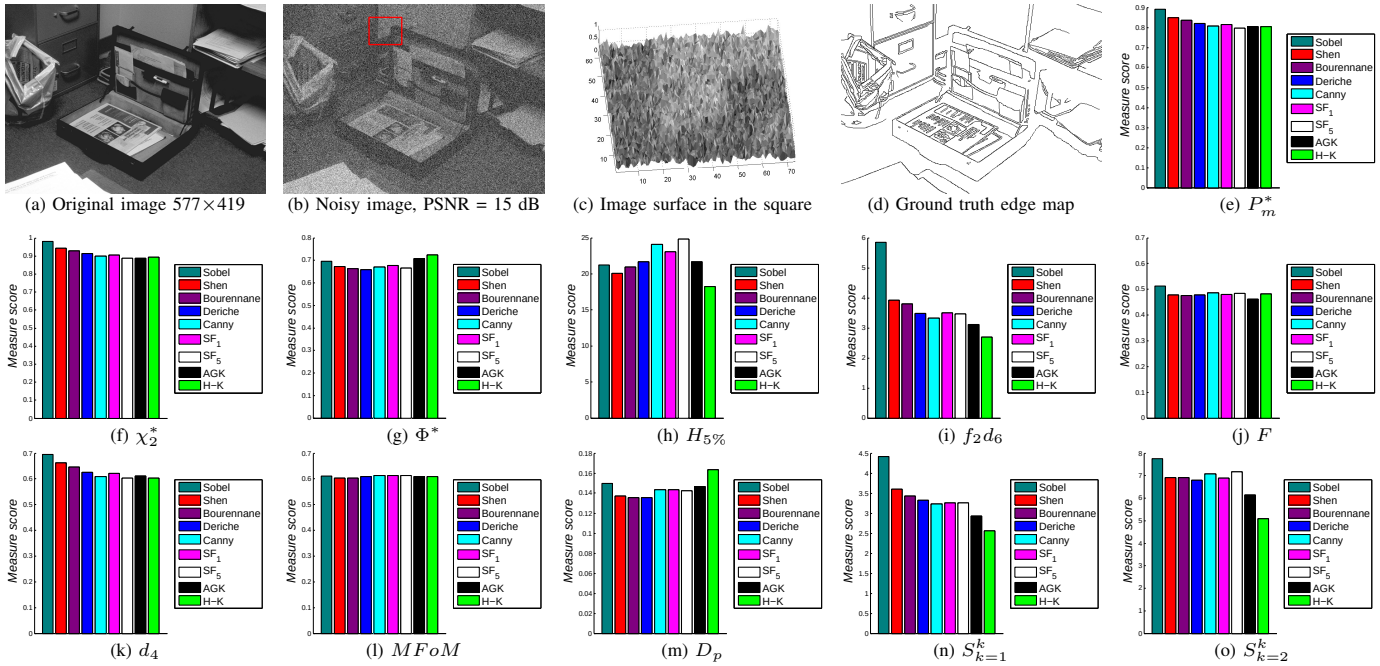
V shows how gradients are computed using these methods. The parameters of the filters are chosen to keep the same spatial support for the derivative information (see Fig. 27(o)). As above, the comparison is the same: the segmented image corresponds to the one having the minimum score for the considered measure.

Box [53] and exponential [50] [6] filters do not delocalize contour points [28] whereas they are sensitive to noise (i.e., addition of FPs). The Deriche [11] and Gaussian filters [8] are less sensitive to noise but suffer from rounding corners and junctions (see [28]) as the oriented filters SF_1 [15], SF_5 [25] and AGK [16], but the more the 2D filter is elongated, the more the segmentation remains robust against noise. Finally, as a compromise, H-K correctly detects contours points having corners and is robust agains noise [35]. Consequently, the scores of the evaluation measures for the first 3 filters must be higher than the three last ones. Furthermore, the ideal segmented image should be visually closer to the ground truth edge image concerning SF_5, AGK and H-K. Thus, Fig. 21 illustrates that F_α gives coherent segmentation and scores even though it is high noisy for images (a)-(f) but the scores are not consistent for the noisy image in Fig. 26(a), see Fig. 27(a). H and Δ^k measures are very sensitive to FPs, therefore, important contours are missing in the segmented images (Figs. 22 and 24). Worse still, concerning the image in Fig. 26(a), the scores of AGK are qualified as the second worst segmentation. The segmented images with respect to FoM remain noisy (Fig. 23) and the scores in Fig. 27(h) penalize both AGK and H-K filters almost like the Sobel filter.

Finally, segmented images using $f_2 d_6$, S^k and Ψ are close to the ground truth edge image and, the more efficient the edge detection filter is, the more the main contours are visible with reasonable FPs and the scores decrease in function of the effectiveness of the filter used. On the contrary, other evaluation measures give either segmented images with high level of FPs, or incoherent scores (bars in Figs. 20 and 27).

TABLE V
GRADIENT MAGNITUDE AND ORIENTATION COMPUTATION FOR A SCALAR IMAGE I WHERE I_θ REPRESENTS THE IMAGE DERIVATIVE USING A FIRST-ORDER FILTER AT THE θ ORIENTATION (IN RADIANS).

Type of operator	Gradient magnitude	Gradient direction
Fixed operator [53], [50], [6], [11], [8]	$\lvert\nabla I\rvert = \sqrt{I_0^2 + I_{\pi/2}^2}$	$\eta = \arctan\left(\frac{I_{\pi/2}}{I_0}\right)$
Oriented Filters [15], [25], [16]	$\lvert\nabla I\rvert = \max_{\theta\in[0,\pi[} \lvert I_\theta\rvert$	$\eta = \arg\max_{\theta\in[0,\pi[} \lvert I_\theta\rvert + \frac{\pi}{2}$
Half Gaussian Kernels [35]	$\lvert\nabla I\rvert = \max_{\theta\in[0,2\pi[} I_\theta - \min_{\theta\in[0,2\pi[} I_\theta$	$\eta = \left(\arg\max_{\theta\in[0,2\pi[} I_\theta + \arg\min_{\theta\in[0,2\pi[} I_\theta\right)/2$

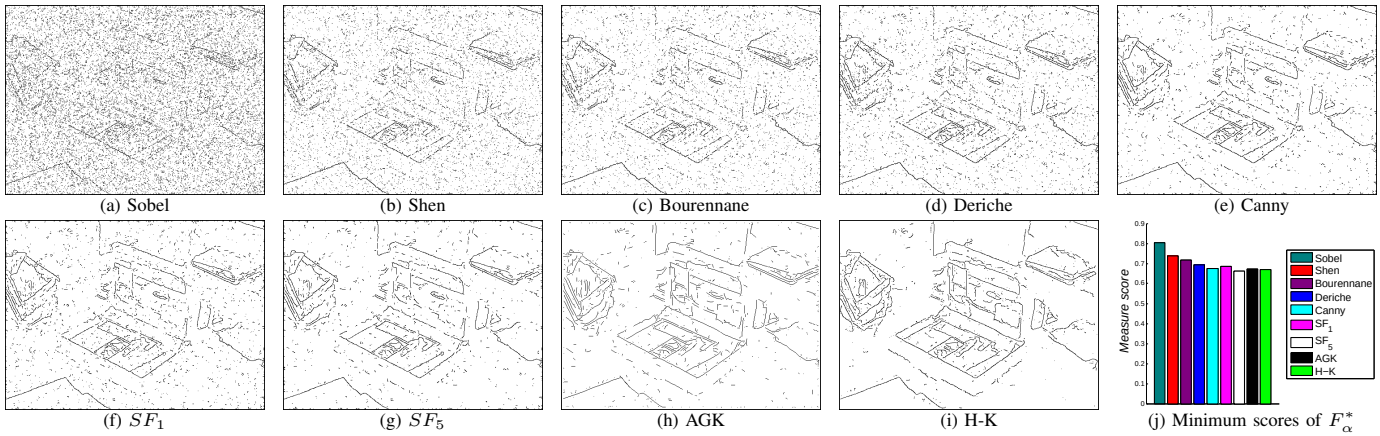


(a) Original image 577×419 (b) Noisy image, PSNR = 15 dB (c) Image surface in the square (d) Ground truth edge map (e) P_m^*

(f) χ_2^* (g) Φ^* (h) $H_{5\%}$ (i) $f_2 d_6$ (j) F

(k) d_4 (l) $MFoM$ (m) D_p (n) $S_{k=1}^k$ (o) $S_{k=2}^k$

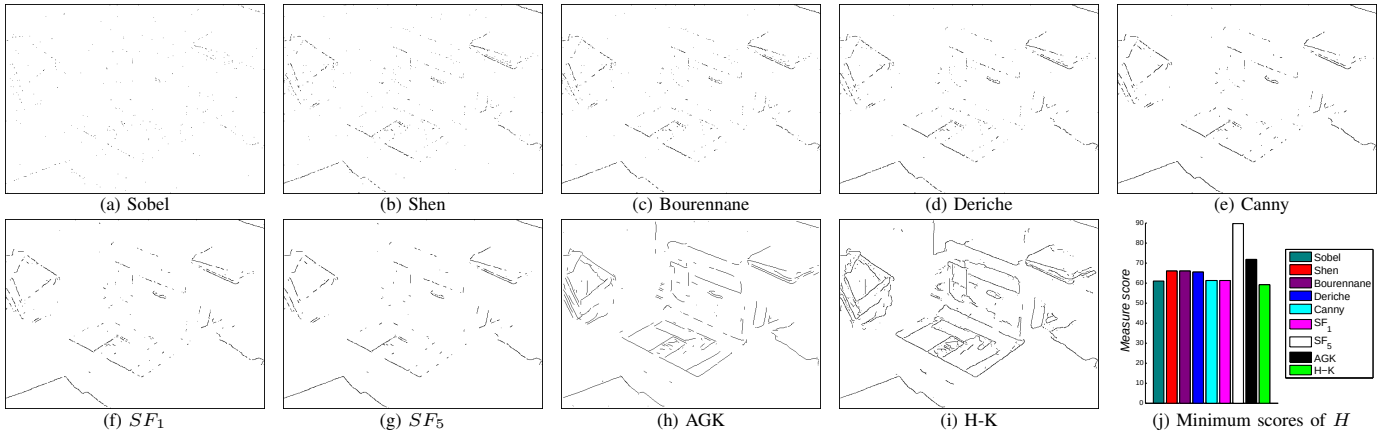
Fig. 20. Comparison of the minimal score for several edge detection evaluation measures concerning a real image and several edge detection methods.



(a) Sobel (b) Shen (c) Bourennane (d) Deriche (e) Canny

(f) SF_1 (g) SF_5 (h) AGK (i) H-K (j) Minimum scores of F_α^*

Fig. 21. Comparison of the minimal score for the F_α measure concerning a real image and several edge detection methods.



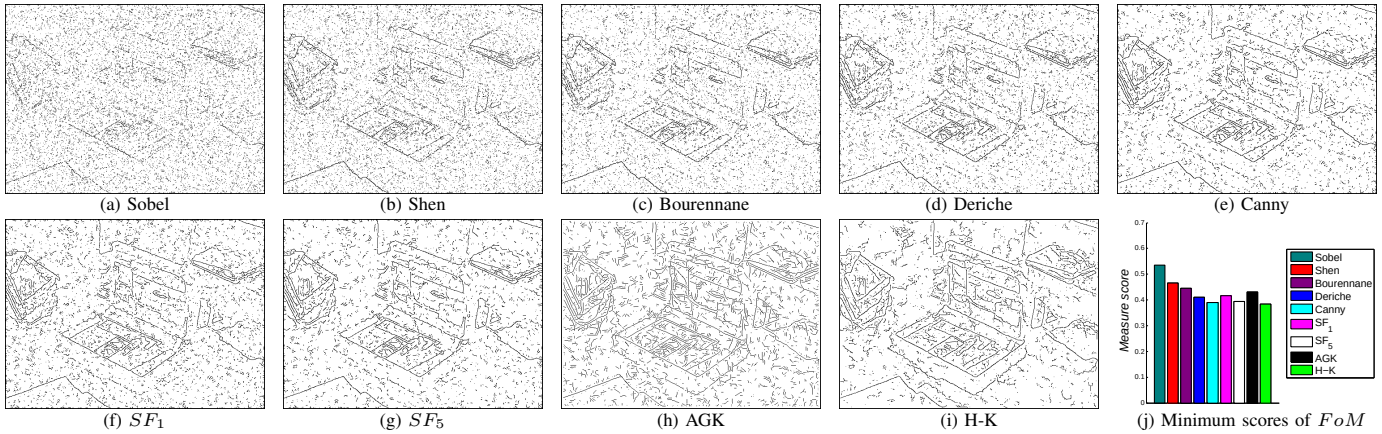
(a) Sobel (b) Shen (c) Bourennane (d) Deriche (e) Canny

(f) SF_1 (g) SF_5 (h) AGK (i) H-K (j) Minimum scores of H

Fig. 22. Comparison of the minimal score for the H measure concerning a real image and several edge detection methods.

V. Conclusion

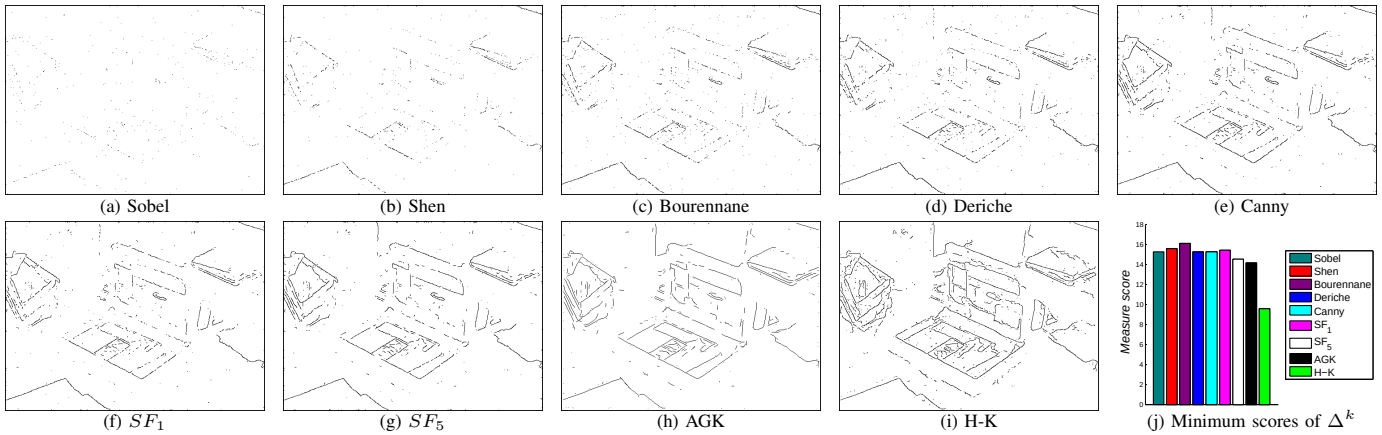
This study presents a review of supervised edge detection assessment methods. A supervised evaluation process estimates scores between two binary images: a ground truth and a candidate edge map. Eight statistical evaluations based on the number of false positive, false negative, true positive or true negative points are detailed. Firstly, examples of the evaluation of contour images prove the need to evaluate an edge detector using assessments involving distance measures. Fifteen distance measures are therefore also evaluated. The



(a) Sobel (b) Shen (c) Bourennane (d) Deriche (e) Canny

(f) SF_1 (g) SF_5 (h) AGK (i) H-K (j) Minimum scores of FoM

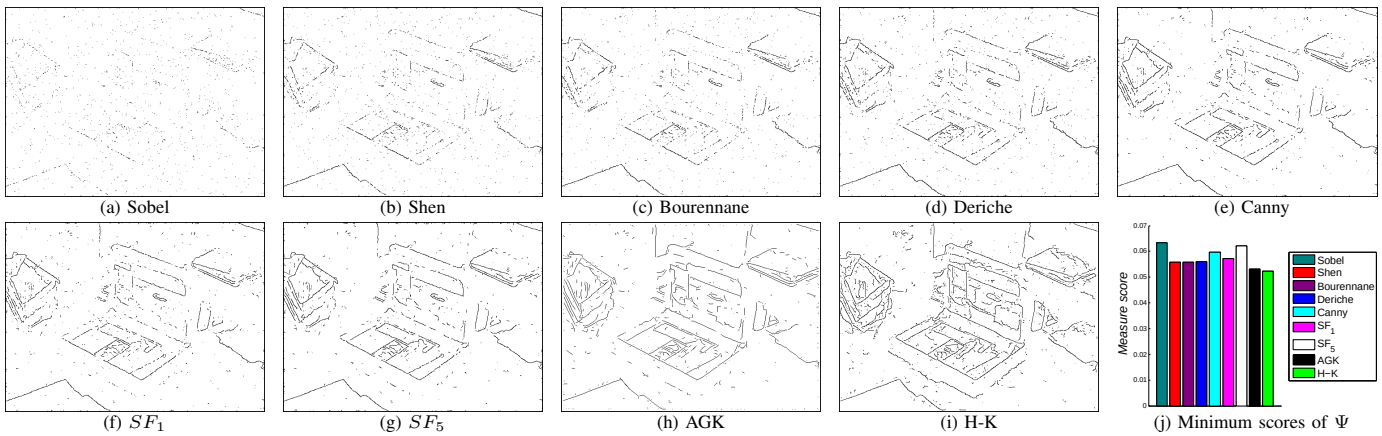
Fig. 23. Comparison of the minimal score for the FoM measure concerning a real image and several edge detection methods.



(a) Sobel (b) Shen (c) Bourennane (d) Deriche (e) Canny

(f) SF_1 (g) SF_5 (h) AGK (i) H-K (j) Minimum scores of Δ^k

Fig. 24. Comparison of the minimal score for the Δ^k measure concerning a real image and several edge detection methods.



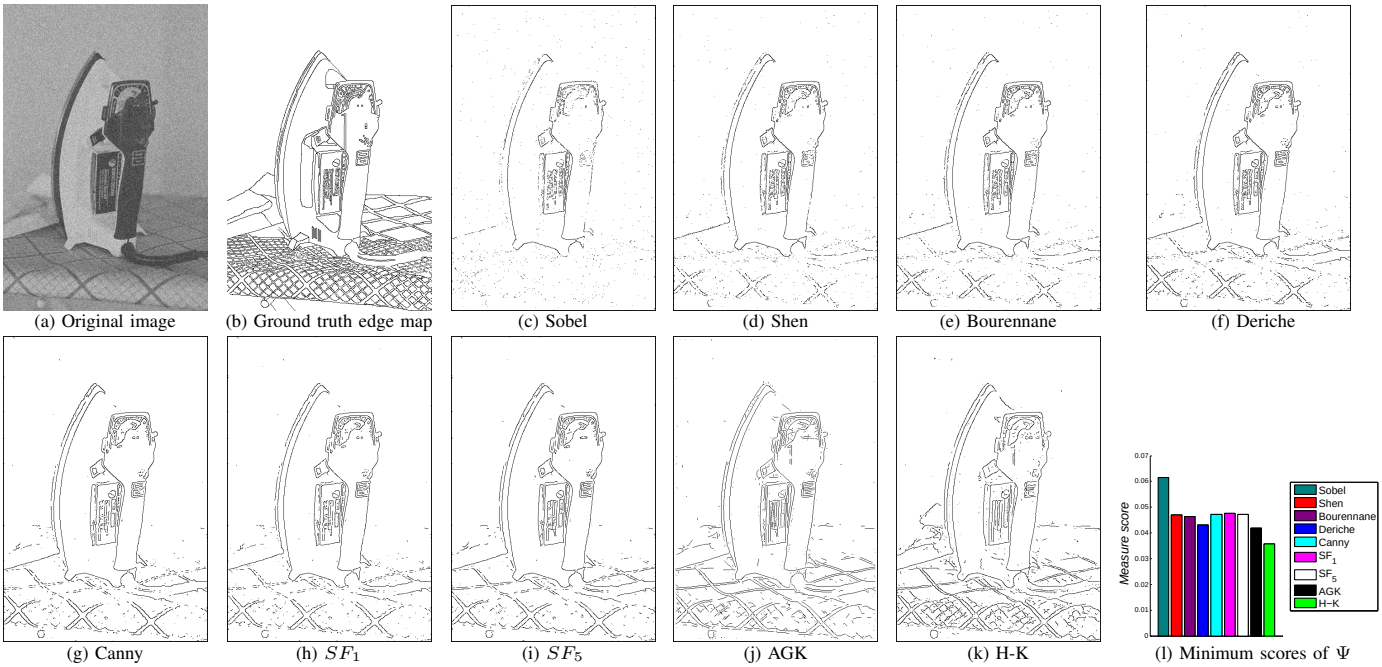
(a) Sobel (b) Shen (c) Bourennane (d) Deriche (e) Canny

(f) SF_1 (g) SF_5 (h) AGK (i) H-K (j) Minimum scores of Ψ

Fig. 25. Comparison of the minimal score for the Ψ measure concerning a real image and several edge detection methods.

latter compute a score depending mainly on the false positive and/or negative distances. By pointing out the drawbacks and advantages of each evaluation measure, in this work a new measure is proposed which takes into account the distance of all the misplaced pixels: false positives and false negatives. Moreover, one of the advantages of the developed method is a function which normalizes the evaluation. Indeed, the score is close to zero concerning good segmentation and increases to one concerning poor edge detection. The influence of the parameters of the new evaluation measure is detailed, theoretically and with concrete cases.

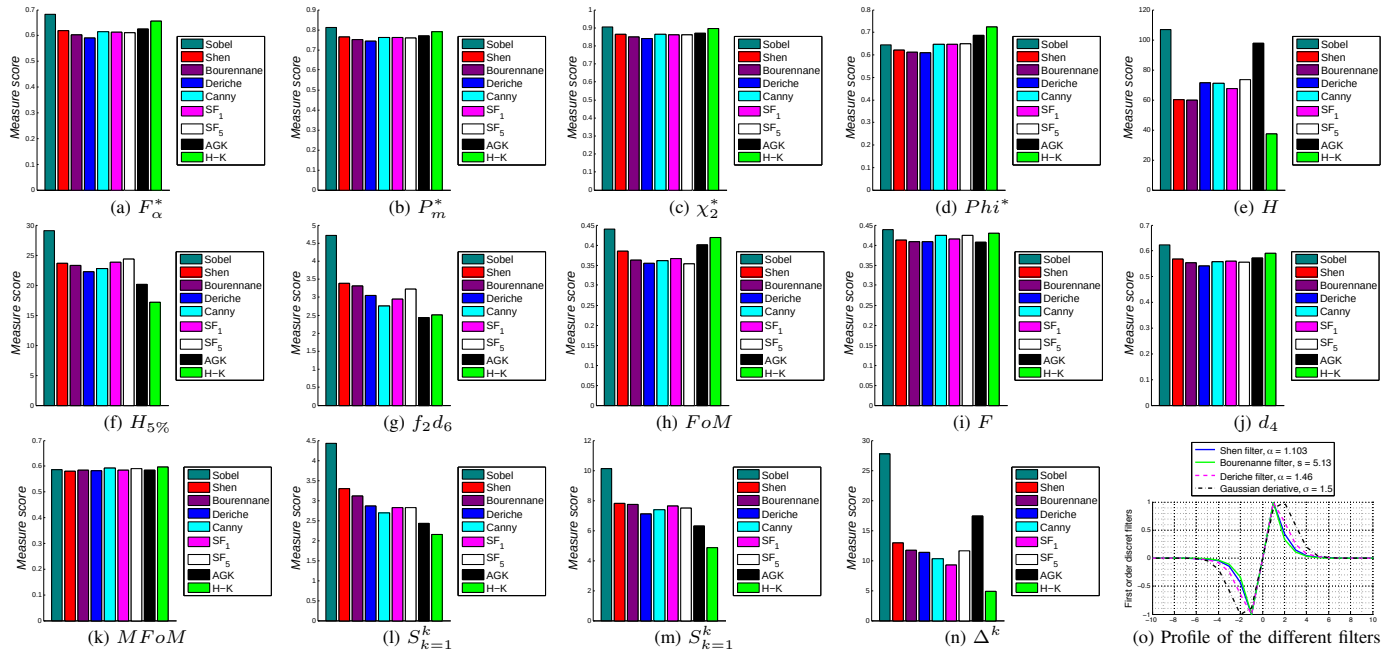
The rest of the paper is dedicated to the experimental results for synthetic and real data. Most of the edge detection evaluation methods are subjected to the following studies: addition of false negative points and/or false positive points, translation of the boundary, remoteness of false positive and false negative contour chains. Finally, as the minimum value of an edge detection evaluation corresponds to the threshold at which the edge detection is the best, the last experiment concerns the minimum value of the measures on image edges compared to the ground truth. This enables the performance of edge detection algorithms to be studied based on filtering



(a) Original image (b) Ground truth edge map (c) Sobel (d) Shen (e) Bourennane (f) Deriche

(g) Canny (h) SF_1 (i) SF_5 (j) AGK (k) H-K (l) Minimum scores of Ψ

Fig. 26. Comparison of the minimal score for the Ψ measure concerning a real image and several edge detection methods.



(a) F_α^* (b) P_m^* (c) χ_2^* (d) Phi^* (e) H

(f) $H_{5\%}$ (g) $f_2 d_6$ (h) FoM (i) F (j) d_4

(k) $MFoM$ (l) $S_{k=1}^k$ (m) $S_{k=1}^k$ (n) Δ^k (o) Profile of the different filters

Fig. 27. Comparison of the minimal score for several edge evaluation measures concerning several edge detection methods of the image in Fig. 26 (a).

through the optimal threshold obtain by the supervised assessment. Therefore, as the proposed function of dissimilarity seems sufficiently robust to evaluate binary edge maps, we plan in a future study to compare more deeply several edge detection algorithms using the optimum threshold computed by the minimum of the evaluation.

REFERENCES

[1] I. E. Abdou and W. K. Pratt, *Quantitative design and evaluation of enhancement/thresholding edge detectors*, Proc. of the IEEE: Vol. 67(5), pp. 753–763, 1979.

[2] A. J. Baddeley, *An error metric for binary images*, Robust Comp. Vision: Quality of Vision Algorithms: pp. 59–78, 1992.

[3] M. Basseville, *Distance measures for signal processing and pattern recognition*, Signal Proc.: Vol. 18(4), pp. 349–369, 1989.

[4] E. Baudrier, F. Nicolier, G. Millon and S. Ruan, *Binary-image comparison with local-dissimilarity quantification*, Pattern Recognition: Vol. 41(5), pp. 1461–1478, 2008.

[5] I. A.G. Boaventura and A. Gonzaga, *Method to evaluate the performance of edge detector*, Brazilian Symposium on Comp. Graphics and Image Proc., 2009.

[6] E. Bourennane, P. Gouton, M. Paindavoine and F. Truchetet, *Generalization of Canny–Deriche filter for detection of noisy exponential edge*, Signal Proc.: Vol. 82(10), pp. 1317–1328, 2002.

[7] K. Bowyer, C. Kranenburg and S. Dougherty, *Edge detector evaluation using empirical ROC curves*, Comp. Vision and Image Understanding: pp. 77–103, 2001.

[8] J. Canny, *A computational approach to edge detection*, IEEE Trans. on Pattern Analysis and Machine Intel.: Vol. 6, pp. 679–698, 1986.

[9] S. Chabrier, H. Laurent, C. Rosenberger, and B. Emile, *Comparative study of contour detection evaluation criteria based on dissimilarity measures*, EURASIP J. on Image and Video Proc.: Vol. 2008, pp. 1–13, 2008.

[10] D. Demigny, *On optimal linear filtering for edge detection*, IEEE Trans. on Image Proc., Vol.11(7), pp.728–737, 2002.

[11] R. Deriche, *Using Canny's criteria to derive a recursively implemented optimal edge detector*, Int. J. Computer Vision, Vol. 1, pp. 167–187, 1987.

[12] E. Deutsch and J. R. Fram, *A quantitative study of the orientation bias of some edge detector schemes*, IEEE Trans. on Computers: Vol. 27(3), pp. 205–213, 1978.

[13] M.P. Dubuisson, A.K. Jain, *A modified Hausdorff distance for object matching*, 12th IAPR Int. Conf. on Pattern Recognition: Vol. 1, pp. 566–568, 1994.

[14] N. L. Fernández-Garcıa, R. Medina-Carnicer, A. and Carmona-Poyato, F. J. Madrid-Cuevas and M. Prieto-Villegas, *Characterization of empirical discrepancy evaluation measures*, Pattern Recognition Letters: Vol. 25(1), 35–47, 2003.

[15] W. T. Freeman and E. H. Adelson, *The design and use of steerable filters*, IEEE TPAMI: Vol. 13, pp. 89–906, 1991.

[16] J.M. Geusebroek, A. Smeulders, and J. van de Weijer, *Fast anisotropic gauss filtering*, ECCV 2002, pp. 99–112, 2002.

[17] J. Gimenez, J. Martinez and A. G. Flesia, *Unsupervised edge map scoring: A statistical complexity approach*, Comp. Vision and Image Understanding: Vol. 122, pp. 131–142, 2014.

[18] A. B. Goumeidane, M. Khamadja, B., Belaroussi, H., Benoit-Cattin and C. Odet, *New discrepancy measures for segmentation evaluation*, IEEE ICIP: Vol. 2, pp. 411–414, 2003.

[19] C. Grigorescu, N. Petkov and M.A. Westenberg, *Contour detection based on nonclassical receptive field inhibition*, IEEE Trans. on Image Proc.: Vol. 12(7), pp. 729–739, 2003.

[20] R. M. Haralick, *Digital step edges from zero crossing of second directional derivatives*, IEEE Trans. on Pattern Analysis and Machine Intel.: Vol. 6(1), pp. 58–68, 1984.

[21] M. D. Heath, S. Sarkar, T. Sanocki, and K. W. Bowyer, *A robust visual method for assessing the relative performance of edge-detection algorithms*, IEEE Trans. on Pattern Analysis and Machine Intel.: Vol. 19(12), pp. 1338–1359, 1997.

[22] B. Hemery, H. Laurent, B. Emile, and C. Rosenberger, *Comparative study of localization metrics for the evaluation of image interpretation systems*, J. of Elec. Imaging: Vol. 19(2), pp. 023017–023017, 2010.

[23] X. Hou, A. Yuille and C. Koch, *Boundary detection benchmarking: Beyond f-measures*, IEEE Conference on Comp. Vision and Pattern Recognition, pp. 2123–2130, 2013.

[24] D. P. Huttenlocher and W. J. Rucklidge, *A multi-resolution technique for comparing images using the Hausdorff distance*, IEEE Trans. on Pattern Analysis and Machine Intel.: Vol. 15(9), pp. 850–863, 1993.

[25] M. Jacob and M. Unser, *Design of steerable filters for feature detection using canny-like criteria*, IEEE TPAMI, Vol. 26(8), pp. 1007–1019, 2004.

[26] R. K. Falah, P. W. Bolon, J. P. Cocquerez, *A Region-Region and Region-Edge Cooperative Approach of Image Segmentation*, IEEE Int. Conf. on Image Proc.: Vol. 3, pp. 470–474, 1994.

[27] U. Köthe, *Reliable low-level image analysis*, Habilitation thesis, 2007.

[28] O. Laligant, F. Truchetet, F. Meriaudeau, *Regularization preserving localization of close edges*, IEEE Signal Proc. Letters, Vol. 14(3), 185–188, 2007.

[29] S. U. Lee, S. Y. Chung, and R. H. Park, *A comparative performance study of several global thresholding techniques for segmentation*, CVGIP, Vol. 52(2), pp. 171–190, 1990.

[30] C. Lopez-Molina, H. Bustince, J. Fernandez, P. Couto and B. De Baets, *A gravitational approach to edge detection based on triangular norms*, Pattern Recognition: Vol. 43(11), pp. 3730–3741, 2010.

[31] C. Lopez-Molina, B. De Baets and H. Bustince, *Quantitative error measures for edge detection*, Pattern Recognition: Vol. 46(4), pp. 1125–1139, 2013.

[32] C. Lopez-Molina, M. Galar, H. Bustince, and B. De Baets, *On the impact of anisotropic diffusion on edge detection*, Pattern Recognition: Vol. 47 (1), pp. 270–281, 2014.

[33] B. Magnier, F., Comby, O., Strauss, J., Triboulet and C. Demonceaux *Highly Specific Pose Estimation with a Catadioptric Omnidirectional Camera*, IEEE Int. Conf. on Imaging Systems and Techniques, pp. 229–233, 2010.

[34] B. Magnier, P. Montesinos, and D. Diep, *Texture Removal by Pixel Classification using a Rotating Filter*, IEEE Int. Conf. on Acoustics, Speech, and Signal Proc., pp. 1097–1100, 2011.

[35] B. Magnier, P. Montesinos and D. Diep, *Fast anisotropic edge detection using gamma correction in color images*, IEEE Int. Symposium on Image and Signal Proc. and Analysis, pp. 212–217, 2011.

[36] B. Magnier, A. Aberkane, P. Borianne, P. Montesinos, and C. Jourdan, *Multi-scale crest line extraction based on half gaussian kernels*, IEEE Int. Conf. on Acoustics, Speech, and Signal Proc., pp. 5105–5109, 2014.

[37] B. Magnier, A. Le and A. Zogo, *A quantitative error measure for the evaluation of roof edge detectors*, IEEE Int. Conf. on Imaging Systems and Techniques, pp. 429–434, 2016.

[38] D. Marr and E. Hildreth, *Theory of Edge Detection*, Proceedings of the Royal Society of London. Series B, Biological Sciences: Vol. 207 (1167), pp. 187–217, 1980.

[39] D. R. Martin, C. C. Fowlkes and J. Malik, *Learning to detect natural image boundaries using local brightness, color, and texture cues*, IEEE Trans. on Pattern Analysis and Machine Intel.: Vol. 26(5), pp. 530–549, 2004.

[40] C. Odet, B. Belaroussi and H. Benoit-Cattin, *Scalable discrepancy measures for segmentation evaluation*, IEEE Int. Conf. on Image Proc.: Vol. 1, pp. 785–788, 2002.

[41] N. Otsu, *A threshold selection method from gray-level histograms*, IEEE Trans. on Systems Man and Cybernetics: Vol. 9, pp. 62-66, 1979.

[42] K. Panetta, C.Gao, S.Agaian,and S.Nercessian, *Non-reference medical image edge map measure*, J. of Bio. Imaging: Vol. 2014, pp. 1–12, 2014.

[43] K. Panetta, C. Gao, S. Agaian and S. Nercessian, *A New Reference-Based Edge Map Quality Measure*, IEEE Trans. on Systems Man and Cybernetics: Systems: Vol. 46(11), pp. 1505–1517, 2016.

[44] G. Papari and N.Petkov, *Edge and line oriented contour detection: State of the art*, Image and Vision Computing: Vol. 29(2) , pp. 79–103, 2011.

[45] J. Paumard, *Robust comparison of binary images*, Pattern Recognition Letters: Vol. 18(10), pp. 1057–1063, 1997.

[46] T. Peli and D. Malah, *A study of edge detection algorithms*, CGIP: Vol. 20(1), pp. 1–21, 1982.

[47] A. J. Pinho and L. B.Almeida, *Edge detection filters based on artificial neural networks. In Int. Conf. on Image Analysis and Proc.*, Springer Berlin Heidelberg, pp. 159–164, 1995.

[48] R. Román-Roldán, J. F. Gómez-Lopera, C. Atae-Allah, J. Martı'nez-Aroza and P.L. Luque-Escamilla, *A measure of quality for evaluating methods of segmentation and edge detection*, Pattern Recognition: Vol. 34(5), 969–980, 2001.

[49] A. Rosenfeld and M. Thurston, *Edge and curve detection for visual scene analysis*, IEEE Trans. on Comp.: Vol. 100(5), pp. 562–569, 1971.

[50] J. Shen and S. Castan, *An optimal linear operator for edge detection*, IEEE CVPR, Vol. 86, pp. 109–114, 1986.

[51] P. Sneath and R. Sokal, *Numerical taxonomy. The principles and practice of numerical classification*, 1973.

[52] K. C. Strasters and J. J. Gerbrands, *Three-dimensional image segmentation using a split, merge and group approach*, Pattern Recognition Letters: Vol. 12(5), pp. 307–325, 1991.

[53] I.E. Sobel, *Camera Models and Machine Perception*, PhD Thesis, Stanford University, 1970.

[54] V. Torre and T.A. Poggio, *On edge detection*, IEEE Trans. on Pattern Analysis and Machine Intel.: Vol. 8, (2), pp. 147–163, 1986.

[55] R. Usamentiaga, D. F. García, C. López, and D. González, *A method for assessment of segmentation success considering uncertainty in the edge positions*, EURASIP J. on Applied Signal Proc., Vol. 2006, pp. 207–207, 2006.

[56] S. Venkatesh and P.L. Rosin, *Dynamic threshold determination by local and global edge evaluation*, Comp. Vision, Graphics, and Image Proc.: Vol. 57(2), pp. 146–160, 1995.

[57] Z. Wang, , A.C. Bovik, H.R. Sheikh and E.P. Simoncelli, *Image Quality Assessment: from Error Visibility to Structural Similarity*, IEEE Trans. on Pattern Analysis and Machine Intel.: Vol. 13(4), pp. 600–612, 2004.

[58] D. L. Wilson, A. J. Baddeley and R. A. Owens, *A new metric for grey-scale image comparison*, Int. J. of Comp. Vision: Vol. 24(1), 5–17, 1997.

[59] W.A. Yasnoff, W. Galbraith and J.W. Bacus, *Error measures for objective assessment of scene segmentation algorithms*, Analytical and Quantitative Cytology: Vol. 1(2), pp. 107–121, 1978.

[60] Y. Yitzhaky and E. Peli, *A method for objective edge detection evaluation and detector parameter selection*, IEEE Trans. on Pattern Analysis and Machine Intel., vol. 25, no. 8, pp. 10271033, 2003.

[61] C. Zhao, W. Shi and Y. Deng, *A new Hausdorff distance for image matching*, Pattern Recognition Letters: Vol. 26(5), pp. 581–586, 2005.