From Contours to Ground Truth: How to Evaluate Edge Detectors by Filtering

Hasan Abdulrahman,
Ecole des Mines d'Alès,
6 avenue de Clavières
30319 Alès, France

Baptiste Magnier
Ecole des Mines d'Alès,
6 avenue de Clavières
30319 Alès, France

Philippe Montesinos
Ecole des Mines d'Alès,
6 avenue de Clavières
30319 Alès, France

ABSTRACT

Edge detection remains a crucial stage in numerous image processing applications. Thus, an edge detection technique needs to be assessed before use it in a computer vision task. As dissimilarity evaluations depend strongly of a ground truth edge map, an inaccurate datum in terms of localization could advantage inaccurate precise edge detectors or/and favor inappropriate a dissimilarity evaluation measure. Hence, in this work, we demonstrate how to label these ground truth data in a semi-automatic way. Moreover, several referenced-based boundary detection evaluations are detailed and applied toward an objective assessment. Thus, each measure is compared by varying the threshold of the thin edges. Indeed, theoretically, the minimum score of the measure corresponds to the best edge map, compared to the ground truth. Finally, experiments on many images using six edge detectors show that the new ground truth database allows an objective comparison of numerous dissimilarity measures.

Keywords

Edge detection, ground truth, supervised evaluation, distance measure, objective evaluation.

1 INTRODUCTION

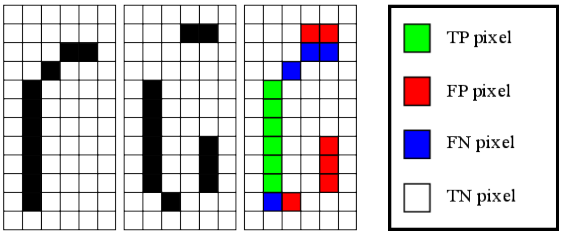
Over the last decades, edge detection remains a crucial role in the computer vision community [30][28][1][42][4][41][8]. This segmentation is considered as a fundamental step in many image processing applications or analysis, pattern recognition, as well as in human vision. Moreover, contours include the most important structures in the image. Typically, edges occur on the boundary between two different regions in an image. In other words, an edge is the boundary between an object and the background or between two different objects.

There exist many different edge detection methods. Nevertheless, an important problem in image processing remains an efficient edge detector comparison and which parameter(s) correspond(s) to the best setting to obtain an accurate edge detection results. Indeed, a robust boundary detection method should create a contour image containing edges at their correct locations with a minimum of misclassified pixels. In order to objectively quantify the performance of an edge detector, a supervised measure computes a similarity/dissimilarity between a segmentation result and a ground truth obtained from synthetic data or a human judgment [2].

In this paper, we detail several edge dissimilarity measures and present how to evaluate filtering edge detection technique involving these considerate measures. In a second time, we demonstrate how to build a new ground truth database which can be used in supervised contour detection evaluation. Indeed, results presented show the importance of the choice of the ground truth. Finally, considering these new ground truth images, results obtained by the measures are exposed.

2 SUPERVISED ERROR MEASURES

To assess an edge detector, 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 original image I. Comparing pixel per pixel G_t and D_c, the first criterion to be assessed



(a) G_t (b) D_c (c) $G_t \cup D_c$ (d) Legend for (c)

Figure 1: Ground truth vs. desired contour. In (b), D_c is contaminated with 6 FPs ans 4 FNs.

Table 1: List of error measures involving only statistics.

Complemented *Performance measure* [3] [4]
$P_m^*(G_t,D_c) = 1 - \frac{TP}{TP+FP+FN}$
Complemented Φ measure [5]
$\Phi^*(G_t,D_c) = 1 - \frac{TPR \cdot TN}{TN+FP}$
Complemented χ^2 measure [6]
$\chi^{2*}(G_t,D_c) = 1 - \frac{TPR-TP-FP}{1-TP-FP} \cdot \frac{TP+FP+FPR}{TP+FP}$
Complemented F_α measure [7]
$F_\alpha^*(G_t,D_c) = 1 - \frac{PREC \cdot TPR}{\alpha \cdot TPR + (1-\alpha)\cdot PREC}$, with $PREC = \frac{TP}{TP+FP}$ and $\alpha \in]0;1]$

is the common presence of edge/non-edge points, as illustrated in Fig. 1. A basic evaluation is compounded from statistics; 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: $FP = |D_c \cap \neg G_t|$,
- **False Negative points (FNs)**, missing boundary points of D_c: $FN = |\neg D_c \cap G_t|$,
- **True Negative points (TNs)**, common non-edge points of G_t and D_c: $TN = |\neg D_c \cap \neg G_t|$.

Computing only FPs and FNs enables a segmentation assessment to be performed [8]. The complemented *Performance measure* P_m^* presented in Table 1 considers directly and simultaneously the three entities TP, FP and FN to assess a binary image [3] [4]. 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) [9] curves or Precision-Recall (PR) [7], involving *True Positive Rates* (TPR) and *False Positive Rates* (FPR): $TPR = \frac{TP}{TP+FN}$ and $FPR = \frac{FP}{FP+TN}$. Derived from TPR and FPR, the three measures Φ, χ^2 and F_α (detailed in Table 1) are frequently used in edge detection assessment. Using the complement of these measures, a score close to 1 indicates a poor segmentation, whereas a value close to 0 a good 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).

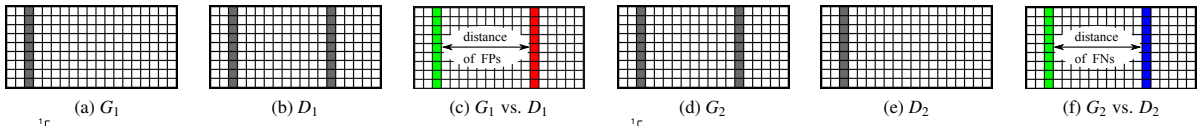
These measures evaluate the comparison of two edge images, pixel per pixel, tending to severely penalize a (even slightly) misplaced contour, as illustrated in Fig. 2 (g) and (h). Thus, to perform an edge evaluation, the assessment should penalize a misplaced edge point proportionally to the distance from its true location.

Table 2: List of normalized error measures compared in this work, with the parameter $\kappa \in]0;1]$.

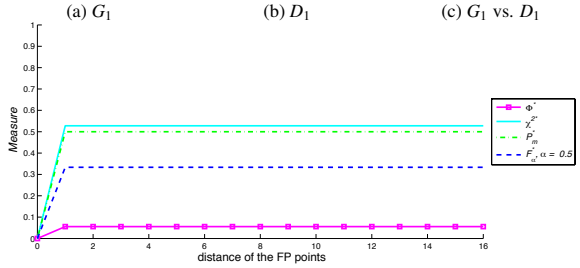
Figure of Merit (*FoM*) [10]
$FoM(G_t,D_c) = 1 - \frac{1}{\max(\lvert G_t\rvert,\lvert D_c\rvert)} \cdot \sum_{p\in D_c} \frac{1}{1+\kappa\cdot d_{G_t}^2(p)}$
FoM of over-segmentation [11]
$FoM_e(G_t,D_c) = 1 - \frac{1}{\max(e^{-FP},FP)} \cdot \sum_{p\in D_c\cap\neg G_t} \frac{1}{1+\kappa\cdot d_{G_t}^2(p)}$
FoM revisited [12]
$F(G_t,D_c) = 1 - \frac{1}{\lvert G_t\cup D_c\rvert} \cdot \sum_{p\in G_t} \frac{1}{1+\kappa\cdot d_{D_c}^2(p)}$
Combination of *FoM* and statistics [13]
$d_4(G_t,D_c) = \frac{1}{2}\cdot\sqrt{S+FoM(G_t,D_c)}$ with $S = \frac{(TP-\max(\lvert G_t\rvert,\lvert D_c\rvert))^2+FN^2+FP^2}{(\max(\lvert G_t\rvert,\lvert D_c\rvert))^2}$
Symmetric Figure of Merit [14]
$SFoM(G_t,D_c) = \frac{1}{2}\cdot FoM(G_t,D_c) + \frac{1}{2}\cdot FoM(D_c,G_t)$
Maximum Figure of Merit [14]
$MFoM(G_t,D_c) = \max(FoM(G_t,D_c), FoM(D_c,G_t))$
Edge map quality measure [15]
$D_p(G_t,D_c) = \frac{1/2}{\lvert I\rvert-\lvert G_t\rvert}\cdot L + \frac{1/2}{\lvert G_t\rvert}\cdot R$ with $L = \sum_{p\in D_c} 1 - \frac{1}{1+\kappa\cdot d_{G_t}^2(p)}$ and $R = \sum_{p\in G_t} 1 - \frac{1}{1+\kappa\cdot d_{G_t\cap D_c}^2(p)}$

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. Tables 2 and 3 review the most relevant measures in the literature. 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 euclidian 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 euclidian distance between p and D_c. On the one hand, some distance measures are specified in the evaluation of over-segmentation (i.e. presence of FPs), like: FoM_e, Υ, D^k, Θ and Γ (see also [26]). On the other hand, Ω measure assesses an edge detection by computing only an under segmentation (i.e. missing ground truth points, see also [26]). Other edge detection evaluation measures consider both FPs and FNs.

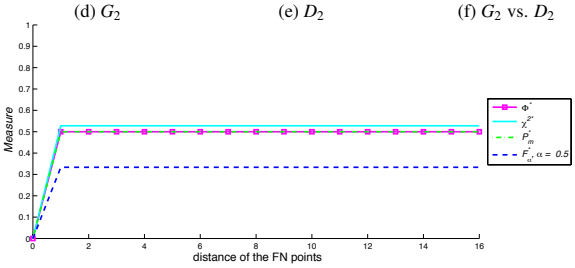
First, to achieve a quantitative index of edge detector performance, one of the most popular descriptors is the Figure of Merit (*FoM*). This distance measure ranges from 0 to 1, where 0 corresponds to a perfect segmentation [10]. Widely utilized for comparing several different segmentation methods, in particular thanks to its normalization criterion, this assessment approach nonetheless suffers from a main drawback. Whenever FNs are created, the distance of FNs ($d_{D_c}(p)$) are not recorded. Indeed, *FoM* can be rewritten as:



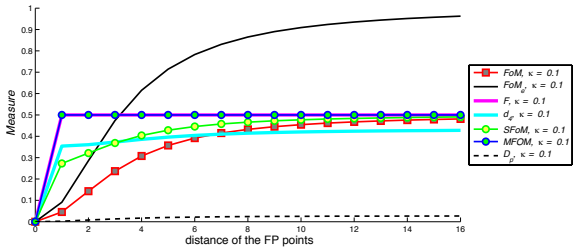
(a) G_1 (b) D_1 (c) G_1 vs. D_1 (d) G_2 (e) D_2 (f) G_2 vs. D_2



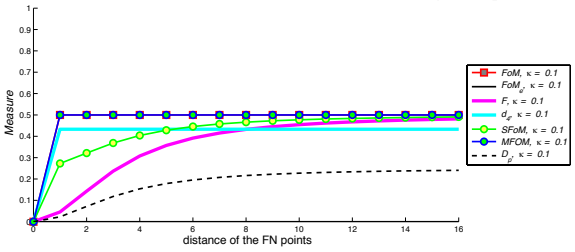
(g) G_1 vs. D_1: Evolution of the confusion matrix-based error assessments in function of the distances of the FPs.



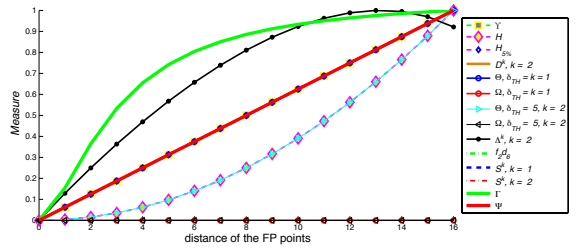
(h) G_2 vs. D_2: Evolution of the confusion matrix-based error assessments in function of the distances of the FNs. Φ^* and P_m^* overlap.



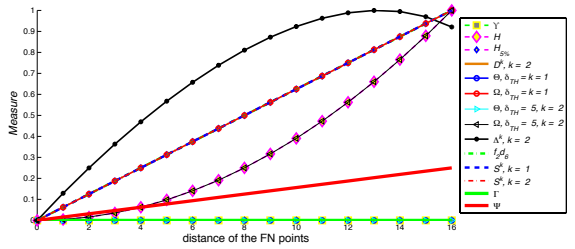
(i) G_1 vs. D_1: Evolution of the normalized dissimilarity measures in function of the distances of the FPs. F and $MFoM$ overlap.



(j) G_2 vs. D_2: Evolution of the normalized dissimilarity measures in function of the distances of the FNs. $FoM_e = 0$. FoM and $MFoM$ overlap.



(k) G_1 vs. D_1: Evolution of the non-normalized dissimilarity measures in function of the distances of the FPs. $\Omega = 0$. H and $\Theta_{\delta_{TH}} = 5$ overlap. Υ, $H_{5\%}$, D^k, $\Theta_{\delta_{TH}} = 1$, $f_2 d_6$, S^k and Ψ overlap.



(l) G_2 vs. D_2: Evolution of the non-normalized dissimilarity measures in function of the distances of the FNs. $\Upsilon = \Theta = D^k = \Gamma = 0$. H and $\Omega_{\delta_{TH}} = 5$ overlap. $H_{5\%}$, $\Omega_{\delta_{TH}} = 1$, $f_2 d_6$ and S^k overlap.

Figure 2: Evolution of dissimilarity measures in function of the the distance of the false positive/negative points. A vertical line of false positive points (b) or false negative points (d) is shifted by a maximum distance of 16 pixels and the measure scores are plotted in function of the displacement of the desired/undesired contour.

$$\begin{aligned} FoM(G_t,D_c) &= 1-\frac{\sum_{p\in D_c\cap G_t}\frac{1}{1+\kappa\cdot d_{G_t}^2(p)}+\sum_{p\in D_c\cap\neg G_t}\frac{1}{1+\kappa\cdot d_{G_t}^2(p)}}{\max(|G_t|,|D_c|)} \\ &= 1-\frac{TP+\sum_{p\in D_c\cap\neg G_t}\frac{1}{1+\kappa\cdot d_{G_t}^2(p)}}{\max(|G_t|,|D_c|)}, \end{aligned} \tag{1}$$

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

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

When $FP=0$, FoM behaves like matrix-based error assessments. Moreover, for $FP>0$, as $\frac{1}{1+\kappa\cdot d_{G_t}^2(p)}<1$, the FoM measure penalizes the over-detection very low compared to the under-detection. The curve in Fig. 2 shows that the penalization of missing points (FNs) becomes higher whereas it is weaker concerning FP. On the contrary, the F measure computes the distances of FNs:

$$F(G_t,D_c)=1-\frac{TP+\sum_{p\in\neg D_c\cap G_t}\frac{1}{1+\kappa\cdot d_{D_c}^2(p)}}{|G_t\cup D_c|}. \tag{3}$$

F behaves inversely to FoM:

$$\begin{cases} \text{if } FP=0: F(G_t,D_c)=1-\frac{|D_c|+\sum_{p\in\neg D_c\cap G_t}\frac{1}{1+\kappa\cdot d_{D_c}^2(p)}}{|G_t|}, \\ \text{if } FN=0: F(G_t,D_c)=1-\frac{|G_t|}{|D_c|}. \end{cases} \tag{4}$$

Also, d_4 measure depends particularly on TP, FP, FN and FoM. Nonetheless, this measure penalizes FNs like the FoM measure, as shown in Fig. 2 (j). $SFoM$ and $MFoM$ take into account both distances of FNS and FPs, so they can compute a global evaluation of a contour image, but as illustrated in Figs. 2 (i) and (j), $MFoM$ does not considers FPs and FNs a the same time, contrary to $SFoM$. Another way to compute a global measure is presented in [15] with the edge map

Table 3: List of non-normalized error measures. In the literature, the most common values are $k = 1$ or $k = 2$.

Yasnoff measure [16]
$\Upsilon(G_t,D_c) = \frac{100}{\lvert I\rvert} \cdot \sqrt{\sum_{p\in D_c} d^2_{G_t}(p)}$
Hausdorff distance [17]
$H(G_t,D_c) = \max\left(\max_{p\in D_c} d_{G_t}(p), \max_{p\in G_t} d_{D_c}(p)\right)$
Distance to G_t [18] [17][19][20]
$D^k(G_t,D_c) = \frac{1}{\lvert D_c\rvert} \cdot \sqrt[k]{\sum_{p\in D_c} d^k_{G_t}(p)}$, $k \in \mathbb{R}^+$, $k = 1$ for [18]
Maximum distance [19]
$f_2d_6(G_t,D_c) = \max\left(\frac{1}{\lvert D_c\rvert}\cdot\sum_{p\in D_c} d_{G_t}(p), \frac{1}{\lvert G_t\rvert}\cdot\sum_{p\in G_t} d_{D_c}(p)\right)$
Oversegmentation [21][22]
$\Theta(G_t,D_c) = \frac{1}{FP}\cdot\sum_{p\in D_c}\left(\frac{d_{G_t}(p)}{\delta_{TH}}\right)^k$, $k \in \mathbb{R}^+$ and $\delta_{TH} \in \mathbb{R}^+_*$ [22], $k = \delta_{TH} = 1$ for [21]
Undersegmentation [21][22]
$\Omega(G_t,D_c) = \frac{1}{FN}\cdot\sum_{p\in G_t}\left(\frac{d_{D_c}(p)}{\delta_{TH}}\right)^k$, $k \in \mathbb{R}^+$ and $\delta_{TH} \in \mathbb{R}^+_*$ [22], $k = \delta_{TH} = 1$ for [21]
Baddeley's Delta Metric [23]
$\Delta^k(G_t,D_c) = \sqrt[k]{\frac{1}{\lvert I\rvert}\cdot\sum_{p\in I}\lvert w(d_{G_t}(p)) - w(d_{D_c}(p))\rvert^k}$, $k \in \mathbb{R}^+$ and a convex function $w : \mathbb{R} \mapsto \mathbb{R}$
Symmetric distance [19][20]
$S^k(G_t,D_c) = \sqrt[k]{\frac{\sum_{p\in D_c} d^k_{G_t}(p)) + \sum_{p\in G_t} d^k_{D_c}(p)}{\lvert D_c\cup G_t\rvert}}$, $k \in \mathbb{R}^+$, $k = 1$ for [19]
Magnier *et al.* measure [24]
$\Gamma(G_t,D_c) = \frac{FP+FN}{\lvert G_t\rvert^2}\cdot\sqrt{\sum_{p\in D_c} d^2_{G_t}(p)}$
Symmetric distance measure [25] [14]
$\Psi(G_t,D_c) = \frac{FP+FN}{\lvert G_t\rvert^2}\cdot\sqrt{\sum_{p\in G_t} d^2_{D_c}(p) + \sum_{p\in D_c} d^2_{G_t}(p)}$

quality measure D_p. 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 more sensitive to FNs than FPs, see Figs. 2 (i) and (j).

A second measure widely computed in matching techniques is represented by the Hausdorff distance H, which measures the mismatch of two sets of points [17]. This max-min distance could be strongly deviated by only one pixel which can be positioned sufficiently far from the pattern. To improve the measure, one idea is to compute H with a proportion of the maximum distances (for example 5% of the values [17]); let us note $H_{5\%}$ this measure. Nevertheless, as pointed out in [19], 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 $H_{n\%}$. To achieve this task, D^k, Υ, Θ and Γ which represent errors of distance only in function of d_{G_t}, they correspond to a measure of over-segmentation (only FPs), as indicated by the curves in Figs. 2 (l) where the curves stagnate at 0. 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). Nevertheless, as concluded in [27], a complete and optimum edge detection evaluation measure should combine assessments of both over- and under-segmentation, as f_2d_6, S^k and Ψ. Also, combining both d_{D_c} and d_{G_t}, Baddeley's Delta Metric (Δ^k) [23] is a measure derived from the Hausdorff distance which is intended to estimate the dissimilarity between each element of two binary images. Finally, curves in Figs. 2 (k) and (l) illustrate that H, $H_{5\%}$, Δ^k, f_2d_6 and S^k behave similarly in function of the FPs or FNs distances. Note that the Ψ measure is more sensitive to the distance of the FPs. The scores of the non-normalized measures in Figs. 2 (k) and (l) are normalized using the following equation for easy visual comparison. Denoting by $f \in [0;+\infty[$ the score vectors of a distance measure such that:

$$\begin{cases} m &= \min(\min(f(G_1,D_1)), \min(f(G_2,D_2))), \\ M &= \max(\max(f(G_1,D_1)), \max(f(G_2,D_2)));\end{cases}$$

then the normalization $\mathcal{N}$ of a measure is computed by:

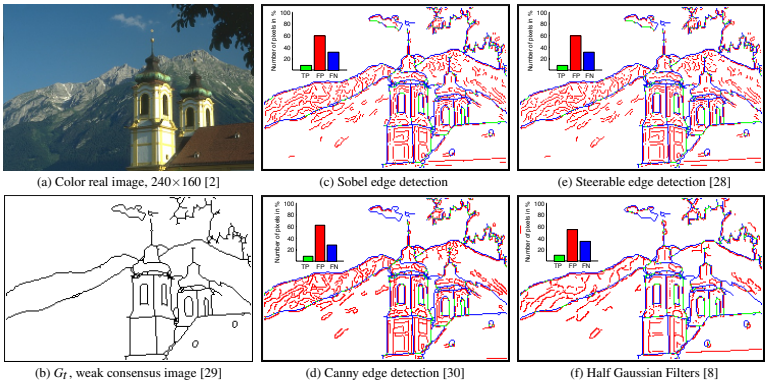
$$\mathcal{N}(f) = \begin{cases} 0 & \text{if } M = m = 0 \\ 1 & \text{if } M = m \neq 0 \\ \frac{f-m}{M-m} & \text{if } M > 1 \text{ and } m \neq 0 \\ f & \text{otherwise.}\end{cases} \quad (5)$$

Other details and behaviors of the different measures are available in [25] and [14]. In the rest of this communication and the supplementary material[1] , the values indicated in the Tables or curves correspond to the true scores of each measure.

3 HOW TO CREATE PRECISE GROUND TRUTH IMAGES? HOW TO EVALUATE A FILTERING TECHNIQUE?

An edge detector is considered as robust when the evaluation score of the dissimilarity with a given G_t is close to 0. Table in Fig.3 reports different assessments for four edge detection methods on a real image (color): Sobel, Canny [30], Steerable Filters (S-F) [28] and Half Gaussian Kernels (H-K) [8]. Only the comparison of D_c with a G_t is studied here. Segmentations are classified together by comparing the scores of the dissimi-

[1] The supplementary material is available at `http://media.wix.com/ugd/c95124_b9338752ae3a4e47852e0fa7bcc58b28.pdf`.



(a) Color real image, 240×160 [2] (c) Sobel edge detection (e) Steerable edge detection [28]

(b) G_t, weak consensus image [29] (d) Canny edge detection [30] (f) Half Gaussian Filters [8]

Meas.	Sobel	Canny	S-F	H-K
P_m^*	0.908	0.908	0.988	0.893
Φ^*	0.766	0.777	0.799	0.779
χ^{2*}	0.987	0.986	0.988	0.979
F_α^*	0.831	0.831	0.838	0.807
FoM	0.399	0.381	0.408	0.344
F	0.705	0.679	0.649	0.6148
Υ	1.007	1.041	1.058	1.025
$\Theta_{\delta_{TH}=1}$	4.323	4.369	4.852	4.577
$\Omega_{\delta_{TH}=1}$	2.783	1.721	1.410	1.640
$H_{5\%}$	19.76	18.45	19.56	20.89
Δ^k	9.454	7.019	8.517	11.13
$S_{k=2}^k$	5.799	5.421	5.626	5.776
Γ	0.499	0.486	0.465	0.393
Ψ	0.584	0.499	0.471	0.402

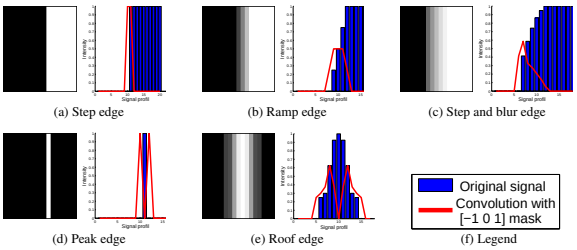
Figure 3: Edge detection after the non-maximum suppression [31] and comparison with a ground truth image.

larity measures and the smallest score for a given measure indicates the best segmentation. Indeed, for example, Sobel corresponds to the best edge detector for Υ, Canny for Δ^k, S-F for Ω and H-F for FoM. However, this assessment suffers from two main drawbacks. Firstly, segmentations are compared using the threshold (voluntary) chosen by the user, this evaluation is very subjective and not reproducible [14]. Secondly, some deficiencies appear in real ground truth contour maps, which could disturb the evaluation of efficient segmentation methods, or, on the contrary, advantage weak/biased edge detectors. Thus, according to the used measure or threshold any detector is classified the first one or the last one.

3.1 Ground truth images

In edge detection assessment, the ground truth is considered as a perfect segmentation. The most common method for ground-truth definition in natural images remains manual labeling by humans [2] [32]. These data sets are not optimal in the context of the definition of low-level segmentation. Firstly, labelers have marked mainly edges of salient objects, whereas equally strong edges in the background or around less important objects are missing. Moreover, errors may be created by human labels (oversights or supplements); indeed, an inaccurate ground truth contour map in terms of localization penalizes precise edge detectors and/or advantages the rough algorithms. There is another problem with the perception. In human perception, images can be ambiguous [33], image structures tend to retain their initial reference (desired shapes) frames, even when rotated or with scale variation(s). In manual segmentation, the perception is influenced by the effects of the particular expectations, the labeler tends to mark easier the contours of a desired object which should be labeled and it influences the result. Finally, in [34], the question is raised concerning the reliability of the datasets regarded as ground truths for novel edge detection methods. Thus, an incomplete ground truth penalizes an algorithm detecting true boundaries and efficient edge detection algorithms obtain between 30% and 40% of errors. Furthermore, when G_t maps are built from a consensus which consists in the combination of several human-labelled images [35][27][29], the deficiencies recalled above remain present. These reasons accentuate the importance of the relevant development of the ground truth labeling.

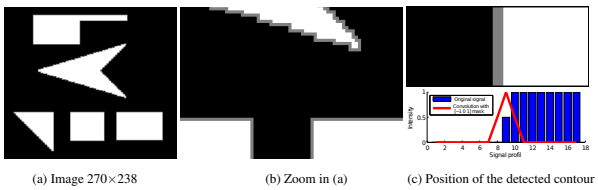
In real digital images, various profile edge types determine contours such as: step, ramp, roof of peaks. Pure step edges are seldom present in real image scenes, but they can be created in synthetic data. As illustrated in Fig. 4, edge positions correspond to the points of the higher gradient magnitude. For a 2D signal, pixels of contours are measured having the higher slope and are localized in the perpendicular direction of the slope of the image function. Considering synthetic data, true edges are positioned between two different colors/gray levels. Nevertheless, the edge position of an object



(a) Step edge (b) Ramp edge (c) Step and blur edge

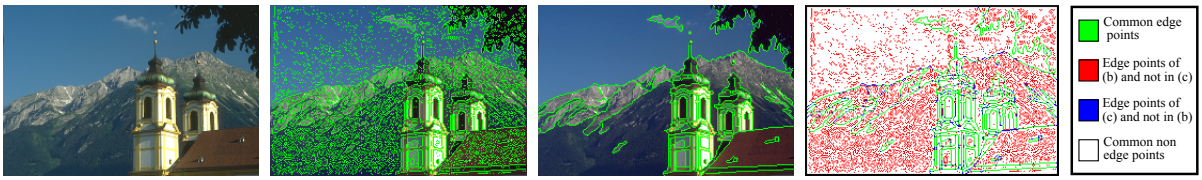
(d) Peak edge (e) Roof edge (f) Legend

Figure 4: The different types of edges and result of a convolution with a $[-1\,0\,1]$ mask (absolute value).



(a) Image 270×238 (b) Zoom in (a) (c) Position of the detected contour

Figure 5: Synthetic data with a 1 pixel width gray around each shape: value of white pixels = 1, values of black pixels = 0, values of gray pixels = 0.5.



(a) Original image (b) Thin edges with $[-1\,0\,1]$ mask (c) Adjustment by hand (d) Image in (b) vs image in (c) (e) Legend

Figure 6: Image of our database are built after an edge detection involving a $[-1\,0\,1]$ mask and concluded by hand.

could be interpreted in different ways: for a vertical step edge, an edge can be located either on the left, or on the right. In Fig. 5, several white shapes are immersed in a black background, creating step edges. To avoid the problem of edge pixel placements, a blur must be voluntary created by adding a 1 pixel width of gray around each shape. Thus, the ground truth corresponds to the points where the slope of the image surface is maximum, i.e. to this gray. These points could be extracted involving odd filters (derivative filters of order 1). In the one hand, a $[-1\,0\,1]$ mask allows to extract the edges at the correct position, i.e. the gray pixels in Fig. 5, contrary to edge detector involving smoothing parameters which delocalize edge positions (especially corners and small objects [1]). In the other hand, using an odd filter, two edges are extracted corresponding to the two boundaries of the roof/peak (see Fig. 4 and [24]); however, human labelers, in majority, indicate only one edge. The new database of contour images issued of real images takes into account all these properties. This paper presents ground truth edge maps which are labeled in a semi-automatic way in order to evaluate the performance of filtering step/ramp edge detectors. Therefore, the motivations to create new ground truth edge images are:

1. To obtain contours accurately localized,
2. To extract edges of the secondary objects or in the background,
3. To exclude boundaries inside noisy/textured regions.

In fact, this new label processes in return to hand made ground truth. Indeed, in a first time, the contours are detected involving the convolution of the image with $[-1\,0\,1]$ vertical and horizontal masks followed by a computation of a gradient magnitude and a suppression of local non-maxima in the gradient direction [31]. Concerning color images, $[-1\,0\,1]$ vertical and horizontal masks are applied to each channel of the image followed by a structure tensor [36]. In a second time, undesirable edges are deleted while missing points are added both by hand. Fig. 6 illustrates the steps to obtain new ground truth images. Using the $[-1\,0\,1]$ mask enables to capture the majority of edge points and corners without deforming small objects, contrary to edge detectors involving Gaussian filters (see for example Fig. 6 in [37]). Moreover, this process enables to detect the good positions of the contours while avoiding the addition of too much imprecise ground truth points, as shown in Fig. 4 and Fig. 5.

3.2 Minimum of the measure

Instead of thresholding manually or automatically [38][39] and then comparing the segmentation of several edge detectors, as in Fig. 8 (c) and (d), the dissimilarity measures are used for an objective assessment. Indeed, the purpose is to compute the minimal value of a dissimilarity measure by varying the threshold Th of the thin edges computed by an edge detector (thin edges are created after the non-maximum suppression of the absolute gradient [31]). 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 [40][27][25]. Fig. 7 illustrates all this process. Since a small threshold leads

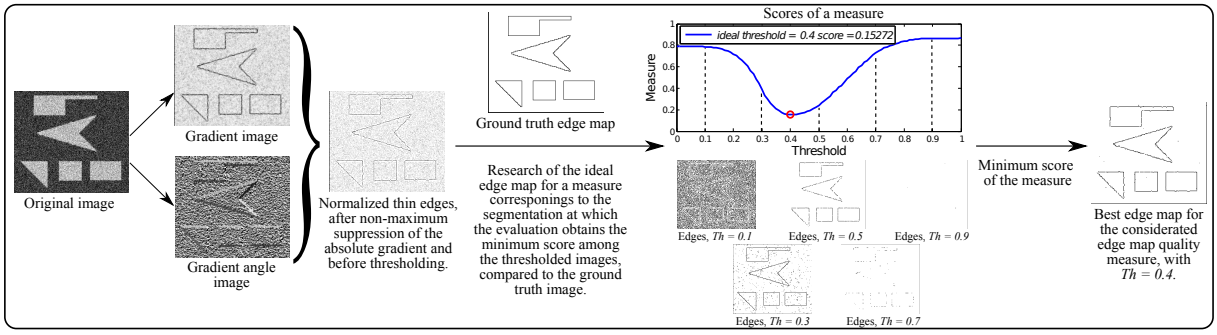
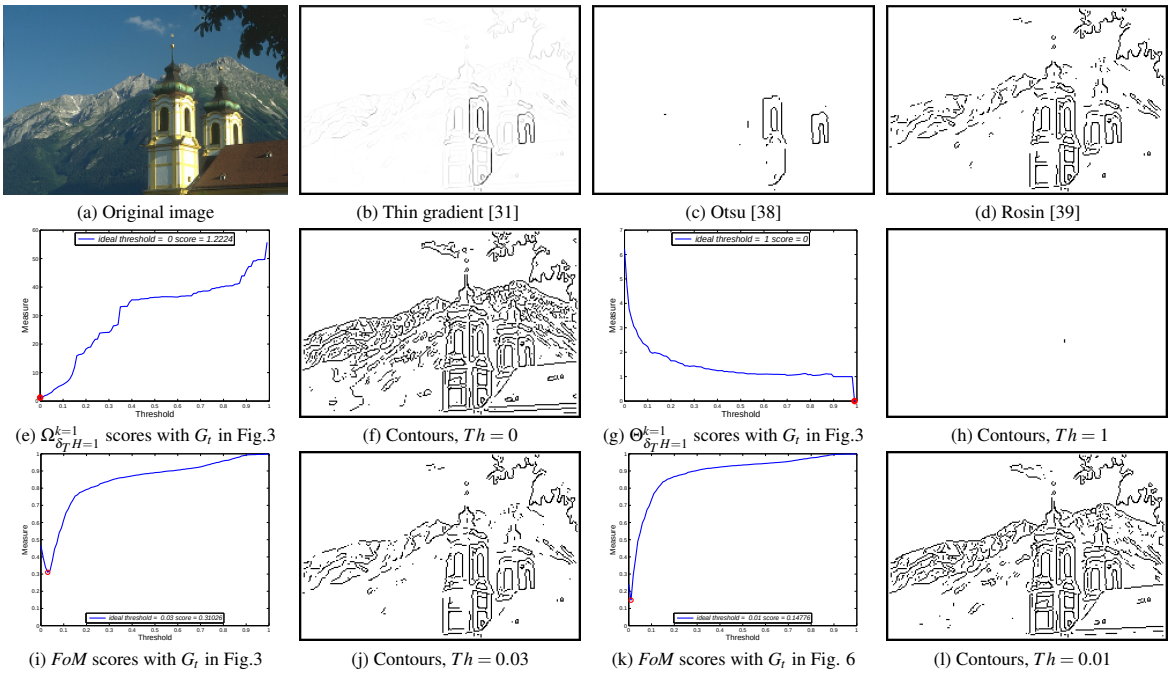


Figure 7: The most relevant edge map for a dissimilarity measure is indicated by its minimum score.



(a) Original image (b) Thin gradient [31] (c) Otsu [38] (d) Rosin [39]

(e) $\Omega^{k=1}_{\delta_T H=1}$ scores with G_t in Fig.3 (f) Contours, $Th = 0$ (g) $\Theta^{k=1}_{\delta_T H=1}$ scores with G_t in Fig.3 (h) Contours, $Th = 1$

(i) *FoM* scores with G_t in Fig.3 (j) Contours, $Th = 0.03$ (k) *FoM* scores with G_t in Fig. 6 (l) Contours, $Th = 0.01$

Figure 8: Scores of the measures depending on the threshold of the thin gradient image [30].

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. As illustrated in Fig. 8 (e) the best score for the under-segmentation evaluation corresponds to $Th = 0$, because false negative points penalize the Ω measure. On the contrary, false positive points penalize over-segmentation dissimilarity measures, as FoM_e, Υ, D^k, Θ and Γ measures, see Fig. 8 (g). Consequently, the best score concerning an over-segmentation measure corresponds to $Th \approx 1$. As G_t are not the same for the evaluation in Fig. 8 (g) and (h), the two curves are different.

4 EXPERIMENTAL RESULTS

The purpose of the experiments presented here is to obtain the best edge map in a supervised way. In order to study the performance of the contour detection evaluation measures, each measure is compared by varying the threshold of the thin edges computed by until six edge detectors: Sobel, Canny [30], Steerable Filters of order 1 (SF_1) [28], Steerable Filters of order 5 (SF_5) [41], Anisotropic Gaussian Kernels (AGK) [42] and Half Gaussian Kernels (HK) [8]. In the one hand, experiments are led on two synthetic noisy images. In the other hand, contour detection evaluations are compared on seven real images where G_t edge maps are labelled by a semi-automatic way (section 3.1). Finally, 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 thin gradient images [30].

Firstly, to evaluate the performances of the dissimilarity measures, the original image in Fig. 5(a) is disturbed with random Gaussian noise and edges are extracted from the noisy images (4dB and 3.3dB, see supplementary material). Generally, the scores of Φ^*, d_4 and D_p measures allow to correctly extract the edges at the price of numerous FPs. Moreover, Δ^k is more sensitive to FPs than the other dissimilarity measures and the best score corresponds to a contour edge map with many discontinuous contours. As pointed out in section 2, concerning the image corrupted by a noise at a level of 4dB, *FoM* penalizes strongly FNs to the detriment of FPs apparitions, and it considers that anisotropic edge detectors are less performant than the Canny edge detector. Other measures classify the Sobel method as the less efficient one and the H-K as the best one.

The second experiment concerns a real image presented in Fig. 6(a); G_t is available in Fig. 6(c). The best edge maps based on minimum of the scores of different measures are presented in Fig. 9. Statistical measures and d_4 consider that Sobel is the best edge detector for this image because edges are well localized. Even though edge maps are different, the scores obtained by *FoM* and F are similar for the different filtering techniques. Oriented kernels, however, are qualified as reliable by distance measures and edge maps corresponding to the minimum scores are less noisy. In the supplementary material are compared the best edge maps obtains with our G_t and G_t of Berkeley segmentation image (Fig. 3(b)). Excepted for Φ^*, d_4 and D_p measures, the best edge map for all the other measures contains many

holes in the contour chains and it is clearly impossible to conclude which edge detector is the most efficient. Table 4 mentions scores involving the two different G_t: by hand made, and semi-automatic. It is important to note that the scores for each measure is smaller concerning G_t built in a semi-automatic way.

Other results presented in the supplementary material show that the minimum scores concerning the distance measures. When objects appear clear, like in image 56 and buildings, most of the measure scores indicate that the edge detectors are equivalent. By contrast, as soon as images contain blur or/and noise, as in image 109 and parkingmeter, the evaluation measures involving error distances considerate that oriented and anisotropic filters produce better-defined contours. Finally, image 109 is a noisy image, however Δ^k and D_p evaluate that Sobel detects better edge, whereas it creates many undesirable contour points, contrary to filtering techniques involving smoothing effects.

Numerous experiments show that $S^k_{k=1\,or\,k=2}$ and Ψ dissimilarity measures are best fitted in the problem of supervised edge evaluation. Indeed, the minimum evaluation scores are coherent and the edge detectors are qualified as best when the filtering technique is adapted to the image structure (blur, noise, small objects). Moreover, the edge map corresponding to the minimum score delimit correctly the object with a majority of continuous contours points without much undesirable points.

5 CONCLUSION

This study presents a review of supervised edge detection assessment methods in details. Moreover, based on the theory of these dissimilarity evaluations, a technique is proposed to evaluate filtering edge detection methods involving the minimum score of the considerate measures. Indeed, to evaluate an edge detection technique, the result which obtains the minimum score of a measure is considerate as the best one and represents an objective evaluation. Theoretically and with the backing of many experiments is demonstrated that the minimum score of the $S^k_{k=1\,or\,k=2}$ and Ψ dissimilarity measures correspond to the best edge quality map evaluations. These two measures take into account both the distances of false positive and false negative points. Many experiments of edge detection on synthetic and real images involving several edge detectors illustrate this conclusion. Experiments show the significance of the ground truth map choice: an inaccurate ground truth contour map in terms of localization penalizes precise edge detectors and/or advantages the rough algorithms. That is the reason why is described in this conversation how to build a new ground truth edge map labelled in semi-automatic way in real images. Firstly, the contours are detected involving the convolution of the image with $[-1\,0\,1]$ masks. Secondly, undesirable edges are removed while missing points are added both by hand, thus a more accuracy ground truth edge map image is built and can be used for supervised contour detection evaluation. By comparison with a real image where contours points are not precisely labelled, experiments illustrate that the new ground truth database allows to evaluate the performance of edge detectors by filtering. Finally, the advantage to compute the minimum score of a measure involving this new ground truth database is that it does not require tuning parameters. For this purpose, we plan in a future study to compare the robustness several edge detection algorithms by adding noise and blur on real images presented in the supplementary material and then using the optimum threshold computed by the minimum of the evaluation.

Table 4: Comparison of scores of dissimilarity measures using a ground truth from [2] (Fig. 3 (b)) image and a constructed ground truth by a semi-automatic way. Contour images and curves for all the measures are available in the supplementary material.

Meas.	Sobel		Canny		SF_1 [28]		AGK [42]		H-K [8]	
	Berkeley G_t	Our G_t	Berkeley G_t	Our G_t	Berkeley G_t	Our G_t	Berkeley G_t	Our G_t	Berkeley G_t	Our G_t
Φ^s	0.738	0.298	0.757	0.430	0.971	0.447	0.813	0.496	0.761	0.504
χ^{2s}	0.979	0.635	0.975	0.725	0.983	0.712	0.982	0.759	0.973	0.502
P^s_m	0.901	0.530	0.901	0.603	0.909	0.594	0.917	0.637	0.893	0.778
F^s_α	0.820	0.360	0.819	0.432	0.834	0.422	0.847	0.468	0.808	0.483
FoM	0.303	0.168	0.310	0.147	0.309	0.164	0.299	0.154	0.277	0.146
F	0.592	0.346	0.579	0.352	0.572	0.310	0.589	0.337	0.589	0.367
d_4	0.675	0.333	0.671	0.379	0.687	0.375	0.695	0.412	0.667	0.424
$SFoM$	0.297	0.145	0.289	0.134	0.270	0.111	0.271	0.119	0.268	0.128
D_P	0.173	0.036	0.184	0.058	0.193	0.056	0.208	0.065	0.183	0.072
H	40.02	29.52	19.41	15.175	18.97	18.02	35.35	14.76	36.87	15.03
$H_{S\%}$	13.72	9.406	11.89	9.142	11.53	6.781	14.18	6.048	14.56	7.165
Δ^k	6.632	4.094	5.039	3.000	4.844	2.462	6.044	2.040	6.562	2.576
f_2d_6	2.851	1.066	2.498	1.294	2.467	0.900	2.625	0.895	2.582	0.983
$S^k_{k=1}$	2.584	1.005	2.315	0.990	2.316	0.877	2.471	0.866	2.432	0.966
$S^k_{k=2}$	4.270	2.323	3.725	2.361	3.690	1.819	4.172	1.667	4.281	2.029
Ψ	0.213	0.041	0.181	0.044	0.173	0.032	0.224	0.032	0.222	0.038

6 ACKNOWLEDGEMENTS

The authors would like to thank the Iraqi Ministry of Higher Education and Scientific Research for funding and supporting this work, and, Philippe Borianne for its helpful comments and suggestions.

7 REFERENCES

[1] D. Ziou and S. Tabbone, "Edge detection techniques: an overview," *Int. J. on Patt. Rec. and Image Anal.*, vol. 8, no. 4, pp. 537–559, 1998.

[2] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *IEEE ICCV*, vol. 2, pp. 416–423, 2001.

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

[4] C. Grigorescu, N. Petkov, and M. Westenberg, "Contour detection based on nonclassical receptive field inhibition," *IEEE TIP*, vol. 12, no. 7, pp. 729–739, 2003.

[5] S. Venkatesh and P. L. Rosin, "Dynamic threshold determination by local and global edge evaluation," *CVGIP*, vol. 57, no. 2, pp. 146–160, 1995.

[6] Y. Yitzhaky and E. Peli, "A method for objective edge detection evaluation and detector parameter selection," *IEEE TPAMI*, vol. 25, no. 8, pp. 1027–1033, 2003.

[7] D. Martin, C. Fowlkes, and J. Malik, "Learning to detect natural image boundaries using local brightness, color, and texture cues," *IEEE TPAMI*, vol. 26, no. 5, pp. 530–549, 2004.

[8] B. Magnier, P. Montesinos, and D. Diep, "Fast anisotropic edge detection using gamma correction in color images," in *IEEE ISPA*, pp. 212–217, 2011.

[9] K. Bowyer, C. Kranenburg, and S. Dougherty, "Edge detector evaluation using empirical roc curves," in *CVIU*, pp. 77–103, 2001.

[10] I. E. Abdou and W. K. Pratt, "Quantitative design and evaluation of enhancement/thresholding edge detectors," *Proc. of the IEEE*, vol. 67, pp. 753–763, 1979.

[11] C. Strasters and J. Gerbrands, "Three-dimensional image segmentation using a split, merge and group approach," *Patt. Rec. Let.*, vol. 12, no. 5, pp. 307–325, 1991.

[12] A. J. Pinho and L. B. Almeida, "Edge detection filters based on artificial neural networks," in *ICIAP*, pp. 159–164, Springer, 1995.

[13] A. G. Boaventura and A. Gonzaga, "Method to evaluate the performance of edge detector," 2009.

[14] H. Abdulrahman, B. Magnier, and P. Montesinos, "A new normalized supervised edge detection evaluation," in *IbPRIA - to appear-*, 2017.

[15] 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, pp. 1505–1517, 2016.

[16] W. Yasnoff, W. Galbraith, and J. Bacus, "Error measures for objective assessment of scene segmentation algorithms.," *Analytical and Quantitative Cytology*, vol. 1, no. 2, pp. 107–121, 1978.

[17] D. Huttenlocher and W. Rucklidge, "A multi-resolution technique for comparing images using the hausdorff distance," in *IEEE CVPR*, pp. 705–706, 1993.

[18] T. Peli and D. Malah, "A study of edge detection algorithms," *CGIP*, vol. 20, no. 1, pp. 1–21, 1982.

[19] M.-P. Dubuisson and A. Jain, "A modified hausdorff distance for object matching," in *IEEE ICPR*, vol. 1, pp. 566–568, 1994.

[20] C. Lopez-Molina, B. De Baets, and H. Bustince, "Quantitative error measures for edge detection," *Patt. Rec.*, vol. 46, no. 4, pp. 1125–1139, 2013.

[21] R. Haralick, "Digital step edges from zero crossing of second directional derivatives," *IEEE TPAMI*, vol. 6, no. 1, pp. 58–68, 1984.

[22] C. Odet, B. Belaroussi, and H. Benoit-Cattin, "Scalable discrepancy measures for segmentation evaluation," in *IEEE ICIP*, vol. 1, pp. 785–788, 2002.

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

[24] B. Magnier, A. Le, and A. Zogo, "A quantitative error measure for the evaluation of roof edge detectors," in *IEEE IST*, pp. 429–434, 2016.

[25] B. Magnier, "Edge detection: a review of dissimilarity evaluations and a proposed normalized measure," *Multimedia Systems for Critical Engineering*, 2017.

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

[27] S. Chabrier, H. Laurent, C. Rosenberger, and B. Emile, "Comparative study of contour detection evaluation criteria based on dissimilarity measures," in *EURASIP J. on Image and Video Proc.*, pp. 1–10, 2008.

[28] W. T. Freeman and E. H. Adelson, "The design and use of steerable filters," *IEEE TPAMI*, vol. 13, pp. 891–906, 1991.

[29] C. Lopez-Molina, B. De Baets, and H. Bustince, "Twofold consensus for boundary detection ground truth," *Knowledge-Based Syst.*, vol. 98, pp. 162–171, 2016.

[30] J. Canny, "A computational approach to edge detection," *IEEE TPAMI*, no. 6, pp. 679–698, 1986.

[31] A. Rosenfeld and M. Thurston, "Edge and curve detection for visual scene analysis," *IEEE Trans. on Computers*, vol. 100, no. 5, pp. 562–569, 1971.

[32] M. Heath, S. Sarkar, T. Sanocki, and K. Bowyer, "A robust visual method for assessing the relative performance of edge-detection algorithms," *IEEE TPAMI*, vol. 19, no. 12, pp. 1338–1359, 1997.

[33] M. Peterson, J. Kihlstrom, P. Rose, and M. Glisky, "Mental images can be ambiguous: Reconstruals and reference-frame reversals," *Memory & Cognition*, vol. 20, no. 2, pp. 107–123, 1992.

[34] X. Hou, A. Yuille, and C. Koch, "Boundary detection benchmarking: Beyond f-measures," in *IEEE CVPR*, pp. 2123–2130, 2013.

[35] N. Fernández-García, A. Carmona-Poyato, R. Medina-Carnicer, and F. Madrid-Cuevas, "Automatic generation of consensus ground truth for the comparison of edge detection techniques," *IVC*, vol. 26, no. 4, pp. 496–511, 2008.

[36] S. Di Zenzo, "A note on the gradient of a multi-image," *CVGIP*, vol. 33, no. 1, pp. 116–125, 1986.

[37] P. Perona and J. Malik, "Detecting and localizing edges composed of steps, peaks and roofs," in *ICCV*, pp. 52–57, IEEE, 1990.

[38] N. Otsu, "A threshold selection method from gray-level histograms," *Automatica*, vol. 11, no. 285-296, pp. 23–27, 1975.

[39] P. L. Rosin, "Unimodal thresholding," *Patt. Rec.*, vol. 34, no. 11, pp. 2083–2096, 2001.

[40] N. Fernández-Garcıa, R. Medina-Carnicer, A. Carmona-Poyato, F. Madrid-Cuevas, and M. Prieto-Villegas, "Characterization of empirical discrepancy evaluation measures," *Patt. Rec. Lett.*, vol. 25, no. 1, pp. 35–47, 2004.

[41] M. Jacob and M. Unser, "Design of steerable filters for feature detection using canny-like criteria," *IEEE TPAMI*, vol. 26, no. 8, pp. 1007–1019, 2004.

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

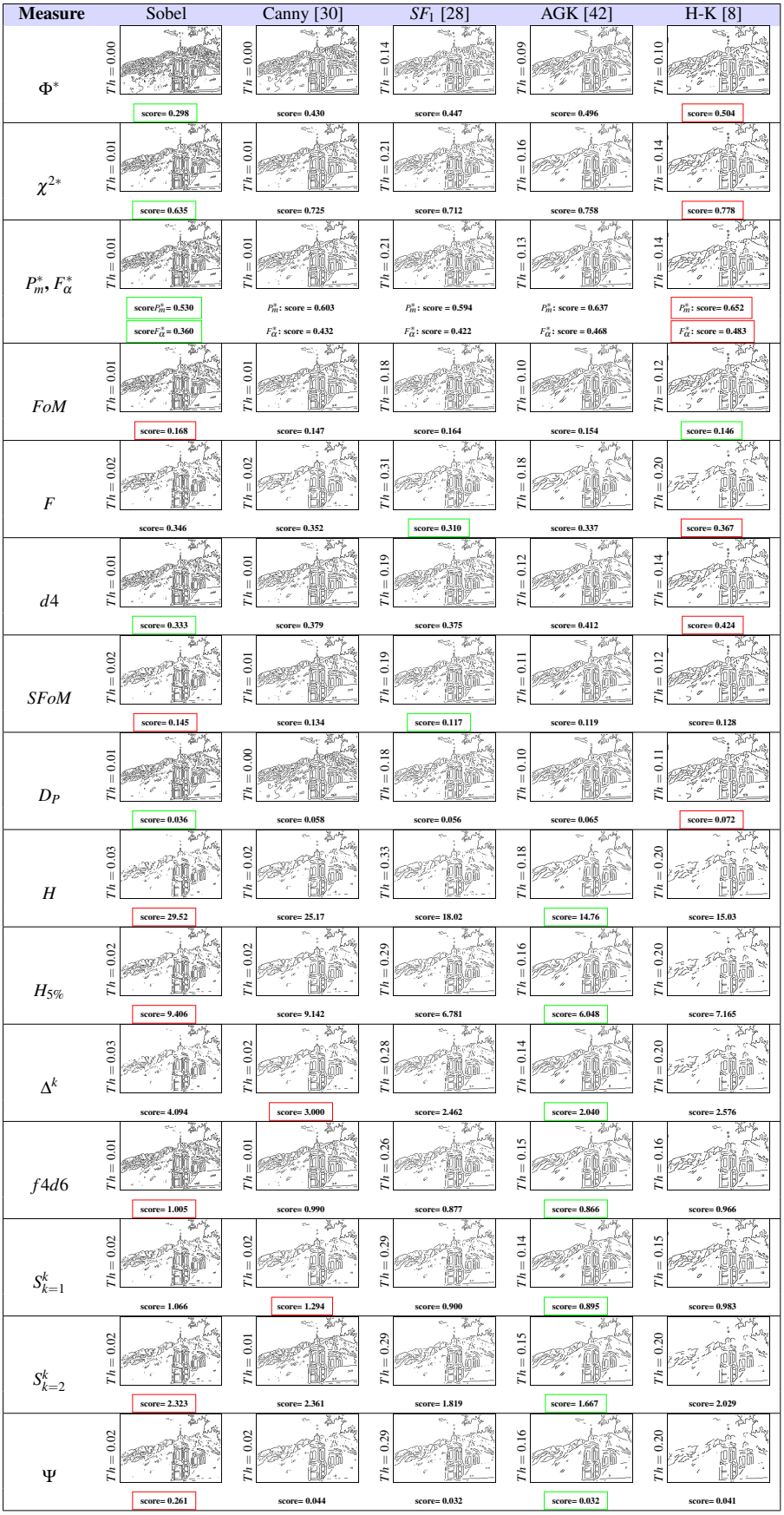


Figure 9: Best edge maps for each dissimilarity measure concerning a real image and G_t in Fig. 6.