A NEW METHODOLOGY FOR EVALUATION OF EDGE DETECTORS

Rodrigo Moreno[1], *Domenec Puig*[1], *Carme Julià*[1], *Miguel Angel Garcia*[2*]

[1] Rovira i Virgili University, Intelligent Robotics and Computer Vision Group
Dept. of Computer Science and Mathematics, Av. Països Catalans 26, 43007 Tarragona, Spain
[2] Autonomous University of Madrid, Dept. of Informatics Engineering
Cra. Colmenar Viejo Km 15, 28049 Madrid, Spain

ABSTRACT

This paper defines a new methodology for evaluating edge detectors through measurements on *edginess maps* instead of on binary edge maps as previous methodologies do. These measurements avoid possible bias introduced by the application-dependent process of generating binary edge maps from edginess maps. The features of completeness, discriminability, precision and robustness, which a general-purpose edge detector must comply with, are introduced. The R, DS, P and FAR-measurements in addition to $PSNR$ applied to the edginess maps are defined to assess the performance of edge detection. The R, DS, P and FAR-measurements can be seen as generalizations of previously proposed measurements on binary edge maps. Well-known and state-of-the-art edge detectors have been compared by means of the new proposed metrics. Results show that it is difficult for an edge detector to comply with all the proposed features.

***Index Terms*—** Edge detection evaluation, completeness, discriminability, precision, robustness.

1. INTRODUCTION

The raw output of a general purpose edge detector can be seen as an edginess map, that is, a map of the probability of every pixel being an edge. Since most applications require binary edge maps instead of edginess maps, post-processing steps, such as hysteresis or thresholding, are then applied to the edginess maps in order to generate such binary maps. However, these post-processing steps are usually application-dependent. For example, edge-based segmentation and image enhancement usually require a different post-processing.

Many measurements have been proposed to assess the performance of edge detectors. The most used ones are the Pratt's Figure of Merit (FOM) [1], the Receiver Operating Characteristic (ROC) curve [2] , Precision vs Recall (PR) curves [3], the F-measure [3], and indirect measurements [4].

The main drawback of these measurements is that they assume that the output of the edge detector is a binary edge map, making it necessary to use application-dependent post-processing steps before the evaluation. Thus, these measurements could be biased by the scope in which the edge detector is being applied, making the results only valid for such a scope. Assessing edge detection before the post-processing steps can avoid those biases, giving a more accurate measure of how good the edge detector is, disregarding the context. Taking this fact into account, this paper proposes a set of desirable features that a general purpose edge detector should comply with and some tools to measure them. These features can be measured directly on the edginess maps, thus avoiding possible bias generated by post-processing steps.

2. QUALITY MEASUREMENT

In general, the edge detection process comprises three steps (see Fig. 1). First, a filtering step is applied to the input image since edge detectors are very sensitive to noise. Second, once the input image is noiseless, edge detectors estimate the likelihood of finding an edge for every pixel. The output of this step is an edginess map. Finally, post-processing is applied to the edginess map in order to obtain a binary edge map. The core of the edge detection algorithms embodies only the first two steps, leaving the post-processing step outside, since this final step is usually application-dependent. In addition, it is not possible to separate the denoising and edginess estimation steps in general, since many algorithms carry out both processes in a unified framework.

The performance of edge detection algorithms can be assessed at three different points of the process, as shown in Fig. 1. Direct measurements at the output of the algorithms are used at the first and second points, while performance at the third point is indirectly assessed by means of measurements of performance of the application in which the edge detector is used. Indirect assessment is based on the assumption that the performance of an edge detector used in the context of a specific application is correlated with the general performance of such an application. Many evaluation methodologies have been proposed to evaluate performance at the second (e.g. [1], [5], [2], [3]) and third points [4], but, to

[*] This research has been partially supported by the Spanish Ministry of Science and Technology under project DPI2007-66556-C03-03, by the Commissioner for Universities and Research of the Department of Innovation, Universities and Companies of Catalonia's Government and by the European Social Fund.

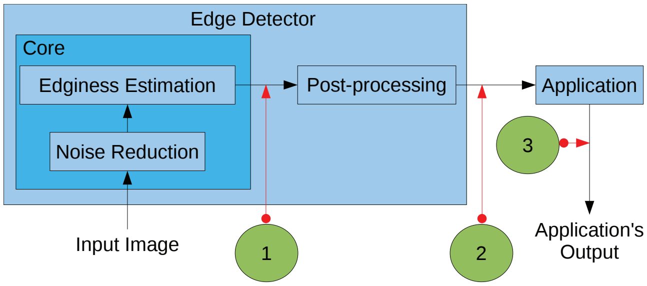

Fig. 1. The edge detection process. The performance of edge detectors can be assessed at the points marked in green.

our knowledge, measuring performance at the first point has not been proposed so far. Despite this, assessing performance at the first point appears advantageous since the core of the edge detectors can be evaluated no matter the context or the applied post-processing. This paper proposes to measure performance at the first point by means of four features: completeness, discriminability, precision and robustness. Without loss of generality, completeness, discriminability and precision can be measured on non-maximum suppressed edginess maps, here referred to as NMSE maps, since the location of edges must be the same, disregarding the strength given to them by the detector. On the other hand, robustness can directly be assessed on the edginess maps. These features are described in the next subsections.

2.1. Completeness

Completeness is referred to as the ability of an edge detector to mark all possible edges of noiseless images. Completeness is a desirable feature of general purpose edge detectors since the decision of whether or not an edge is relevant only depends on the application. Thus, an edge detector reduces its scope when it decides to discard edges. Despite that, most edge detectors usually opt for decreasing their scope for the sake of improving their performance in other features, such as discriminability or robustness.

Complete ground-truths with all the possible edges must be used to measure completeness. Unfortunately, that kind of ground-truths are not usually available. Thus, recall, the ground-truth dependent counterpart of completeness, can be used to give partial estimations of completeness. Let d_i be the distance between the i-th pixel in the ground-truth and its corresponding matching pixel in the NMSE map or infinity if such a matching pixel does not exist, m and n be the number of marked pixels in the ground-truth and in the NMSE map respectively, and $\phi(\cdot)$ be a radial decaying function in the range from zero to one. The R-measurement, given by $R = \frac{1}{m}\sum_{i=1}^{m}\phi(d_i)$, can be used to estimate recall. An issue related to measuring recall when $n > m$ is the fact that every edge detector generates a different number of edges and this can give advantage to detectors that generate higher amounts of edges, since d_i tends to be reduced when n increases. These bias can be eliminated by taking the same number of detected edges for all the edge detectors to be compared. This can be done by taking the n' strongest detected edges taken from the NMSE maps. R vs. n' plots can also be used to analyze the evolution of R.

2.2. Discriminability

Discriminability is referred to as the ability of an edge detector to discriminate between important and not important edges. This feature is application-dependent since importance can only be assessed in a specific scope. For example, the discriminability of an edge detector could be high when applied to image enhancing or low when applied to edge-based segmentation. Discriminability is one of the most desirable features of edge detectors since low levels of discriminability make it necessary to use more sophisticated post-processing algorithms that can partially fix the drawbacks of the edge detector. Thus, global thresholding (the simplest post-processing algorithm) could be used for edge detectors with maximum discriminability. Discriminability can be measured related to a specific ground-truth through the DS-measurement, the difference between the weighted mean edginess of the pixels that match the ground-truth and the weighted mean edginess of the pixels that do not match it. Let e_i be the edginess at pixel i of the NMSE map, usually in the 0 to 255 range. The DS-measurement is given by:

$$DS = \frac{\sum_{i=1}^{n} e_i\phi(d_i)}{\sum_{i=1}^{n}\phi(d_i)} - \frac{\sum_{i=1}^{n} e_i(1-\phi(d_i))}{\sum_{i=1}^{n} 1-\phi(d_i)}. \quad (1)$$

2.3. Precision

Precision measures the ability of an edge detector to mark edges as close as possible to real edges. Precision is mandatory in edge detection, since the performance of applications in which the detectors are used depends on this feature. Unlike discriminability, precision is in essence an application-independent feature. However, in practice, application-independent measures of precision are difficult to obtain since complete ground-truths are required. Thus, only precision measurements related to specific ground-truths can be obtained. Ideally, all edges of the ground-truth should be found at distance zero in the NMSE map. However, if hand-made ground-truths are used, it is necessary to take into account that those ground-truths are not precise, since some pixels can appear misplaced due to human errors. Despite that, those ground-truths can still be used to compare edge detectors, since all edge detectors are equally affected by these errors. Let $\overline{D}$ be the mean distance from pixels of the ground-truth that match the NMSE map. The P-measurement, given by $P = \phi(\overline{D})$, can be used to estimate precision.

A feature related to precision is the false alarm rejection feature, which represents the ability of edge detectors not to detect edges in flat regions. The FAR-measurement, given by $FAR = \frac{1}{n}\sum_{i=1}^{n}\phi(d_i)$, can be used as a numeric ground-truth dependent estimation of false alarm rejection. Similarly to the R-measurement, the n' strongest detected edges from the NMSE map must be selected before the calculations of the P and FAR-measurements in order to avoid biases related to n when $n > m$. Thus, plots of P vs. n' and FAR vs. n' can also be used to evaluate the evolution of the P and FAR-measurements.

2.4. Robustness

Robustness measures the ability of an edge detector to reject noise. Thus, an ideal robust edge detector should produce the same output for both noisy and noiseless images. Robustness is one of the most difficult features to comply with since edge detection is essentially a derivative operation which is usually very sensitive to noise. In fact, most edge detectors include filtering steps in order to improve their robustness. However, most of those filters mistakenly eliminate important details treating them as noise, reducing thus the completeness and recall features of the detector. Despite that, robustness is usually preferred to completeness.

Since edginess maps can be modeled by means of grey-scale images, metrics of image fidelity can be used to measure robustness. The peak signal to noise ratio (PSNR) is the most widely used metric of image fidelity. Although PSNR is not suitable to measure precision [6], it is appropriate to measure robustness. The edge detector is applied to both the noiseless and the noisy version of the same image. The PSNR between both outputs is calculated in order to measure the difference between them. Unlike the above described measurements, it is not necessary to use ground-truths or to apply non-maximum suppression to the edginess maps before computing PSNR. Let r and c be the dimensions of the edginess maps, and e_{ij} and e'_{ij} be the edginess at location i, j of the noisy and noiseless images respectively. The PSNR is given by $\text{PSNR} = 10\ \log_{10}(255^2/\text{MSE})$, with

$$\text{MSE} = \frac{1}{r\,c}\sum_{i=1}^{r}\sum_{j=1}^{c}\left(e_{ij} - e'_{ij})\right)^2 . \qquad (2)$$

3. RELATIONS TO OTHER ASSESSMENT METRICS

The R, DS, P and FAR-measurements can be seen as generalizations of the Pratt's FOM [1] by selecting $\phi(x) = 1/(1+\alpha x^2)$, with α being a constant. Related metrics such as the Pratt's FOM [1] and the pixel correspondence metric [6] are unable to measure recall, precision and false alarm rejection separately. PR curves [3], ROC curves [2] and the F-measure [3] are based on the measurement of three features of edge detectors: precision, recall, and false alarm rejection on binary edge maps. All of them are based on the classification of the detected edges into four groups: true positives, true negatives, false positives and false negatives. Unfortunately, a correct classification of edges is only possible for complete ground-truths, and only incomplete, application-dependent ground-truths are usually available for natural images. Thus, for example, a detected edge which does not match a specific ground-truth could be misclassified as a false positive, since it still could match an edge if the complete ground-truth is used instead. Similarly, a pixel that is not marked both by the detector and by the specific ground-truth could be misclassified as a true negative, since an additional true edge still could appear at the pixel's location in the complete ground-truth. The proposed measurements can evaluate these three features more effectively since they do not require to classify edges.

Table 1. Performance measurements for GT1 and GT2.

Method	R		DS		P		FAR	
	GT1	GT2	GT1	GT2	GT1	GT2	GT1	GT2
Compass	0.57	0.61	8.62	41.10	0.93	0.71	0.89	0.57
LGC	0.15	0.40	4.46	**44.33**	**0.96**	0.85	**0.93**	**0.78**
LoG	0.44	0.68	9.45	38.81	0.93	0.63	0.90	0.51
Sobel	**0.84**	**0.75**	**9.89**	34.07	0.94	**0.88**	0.84	0.74

4. RESULTS

A total of fifteen images from the Berkeley segmentation data set [7] have been used in the experiments. In addition to the Laplacian of Gaussians (LoG) and the Sobel detectors, the methods proposed by Maire *et al.* [8], referred to as the LGC method, and by Ruzon and Tomasi [9], referred to as the Compass method, have been used in the comparisons, since they are considered to represent the state-of-the-art in edge detection. The Compass and LoG algorithms have been applied with $\sigma = 2$, while the default parameters of the LGC method have been used. Three ground-truths have been considered in the experiments: the NMSE map generated by the Prewitt's edge detector, a computer generated consensus ground-truth [10], which could be used in the image enhancement application, and the hand-made ground-truth of the Berkeley segmentation data set [7], whose validity has only been proven in segmentation related applications. We will refer to those ground-truths as GT1, GT2 and GT3 respectively. We used the function $\phi(x) = 1/(1+\alpha x^2)$ with $\alpha = 1/9$ as suggested in [1]. We matched every detected edge to its closest pixel in the ground-truth, restricting up to one match for every ground-truth pixel. However, more sophisticated matching algorithms could be used (e.g. [3], [6]). Gaussian noise with different standard deviations has been added to the images for the robustness analysis. Table 1 shows the proposed performance measurements for GT1 and GT2. Figure 2 shows the evolution of the proposed performance measurements for GT3 with n'. Evolution plots are not necessary for GT1 and GT2 since $m \geq n$ for them.

As for GT1, all the tested algorithms have a good precision and false alarm rejection but a poor discriminability. Only Sobel has a good performance in recall. These results were expected since $m \geq n$. As for GT2, LGC is the best algorithm in discriminability and false alarm rejec-

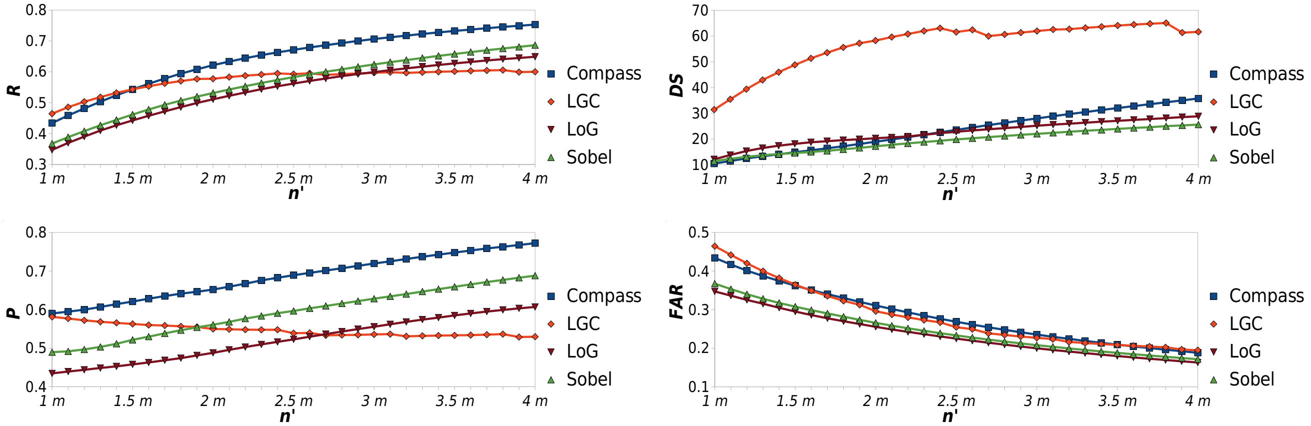

Fig. 2. Performance measurements for GT3: top left: R vs. n' (recall); top right: DS vs. n' (discriminability); bottom left: P vs. n' (precision); bottom right: FAR vs. n' (false alarm rejection). n' is given in terms of m.

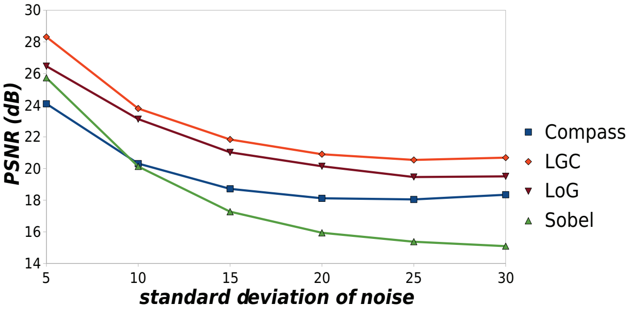

Fig. 3. Robustness analysis.

tion. However, LGC shows the worst value of recall. Sobel is the best in precision and recall and has a competitive performance in FAR and DS. As for GT3, Fig. 2 shows that the Compass detector has the best evolution for the R and P-measurements, the LCG is the best for the DS-measurement and both of them have a similar performance in FAR. The performance of the Sobel and LoG detectors is the worst, but it increases with n' even surpassing in R and P the LGC for high values of n'. As for the robustness analysis (see Fig. 3), LGC appears to be the most robust algorithm with around 1 dB above the LoG. The Sobel and Compass detectors are more sensitive to noise.

5. CONCLUSIONS

A set of metrics for edge detection evaluation has been presented. Those metrics aim at measuring the features of completeness, discriminability, precision and robustness on edginess maps. The main advantage of these proposed measures is that they are directly applied to the edginess maps, avoiding possible bias generated by post-processing steps. Results show that it is difficult for an edge detector to have a good performance for all the metrics. Thus, an edge detector should be chosen for every particular application depending on the required performance.

6. REFERENCES

[1] W. Pratt, *Digital Image Processing: PIKS Scientific Inside*, Wiley-Interscience, fourth edition, 2007.

[2] K. Bowyer, C. Kranenburg, and S. Dougherty, “Edge detector evaluation using empirical ROC curves,” *Comput. Vis. and Image Underst.*, vol. 84, no. 1, pp. 77–103, 2001.

[3] D. Martin, C. Fowlkes, and J. Malik, “Learning to detect natural image boundaries using local brightness, color and texture cues,” *IEEE Trans. PAMI*, vol. 26, no. 1, pp. 530–549, 2004.

[4] M. Shin, D. Goldgof, K. Bowyer, and S. Nikiforou, “Comparison of edge detection algorithms using a structure from motion task,” *IEEE Trans. on Syst., Man and Cybern. - B*, vol. 31, no. 4, pp. 589–601, 2001.

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

[6] M. Prieto and A. Allen, “A similarity metric for edge images,” *IEEE Trans. PAMI*, vol. 25, no. 10, pp. 1265–1273, 2003.

[7] 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 *Proc. ICCV*, 2001, pp. II:416–423.

[8] M. Maire, P. Arbelaez, C. Fowlkes, and J. Malik, “Using contours to detect and localize junctions in natural images,” in *Proc. CVPR*, 2008, pp. 1–8.

[9] M. Ruzon and C. Tomasi, “Edge, junction, and corner detection using color distributions,” *IEEE Trans. PAMI*, vol. 23, no. 11, pp. 1281–1295, 2001.

[10] 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,” *Image and Vis. Comput.*, vol. 26, no. 4, pp. 496–511, 2008.