Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics
San Antonio, TX, USA - October 2009

A Non-Reference Measure for Objective Edge Map Evaluation

Shahan Nercessian, Karen Panetta
Department of Electrical and Computer Engineering
Tufts University
Medford, MA, USA
shahan.nercessian@tufts.edu, karen@eecs.tufts.edu

Sos Agaian
Department of Electrical and Computer Engineering
University of Texas at San Antonio
San Antonio, TX, USA
sagaian@utsa.edu

***Abstract*— Edge detection has been used extensively as a pre-processing step for many computer vision tasks. Due to its importance in image processing and the highly subjective nature of human evaluation and visual comparison of edge detectors, it is desirable to formulate objective edge map evaluation measures. One would like to use such a measure to make comparisons of results using the same edge detector with different parameters as well as to make comparisons of results using different edge detectors. Reconstruction-based measures have the clear advantage that they effectively incorporate original image data. In this paper, a general model for reconstruction-based measures is established in order to alleviate the shortcomings of the reconstruction-based measures, followed by the formulation of a new non-reference measure for objective edge map evaluation. Experimental results illustrate the effectiveness of the new measure both as a means of selecting optimal edge detector parameters and as a means of determining the relative performance of edge detectors for a given image.**

***Keywords*—Edge detection, objective evaluation, performance measures**

I. INTRODUCTION

Edge detection plays a great role in image processing and computer vision tasks as it determines the structure of objects in images. Therefore, it can aid processes by substantially reducing the amount of information to be processed [1]. As a result, edge detection has served as a basis for many feature extraction, object detection, object recognition, image enhancement, and image segmentation algorithms, and has been used extensively for remote sensing, security systems, handwriting analysis, and biomedical applications [2, 3, 4]. As edge detection has been used as a pre-processing step for so many algorithms in image processing, effective and objective edge map evaluation measures must be developed in order to assess edge detector performance. Objective edge map evaluation measures have many important uses. Obviously, after the development of so many different edge detection algorithms over the years, there should be an objective way of determining which edge detection algorithm generally performs the best. Secondly, an objective edge map evaluation measure can select the tunable parameters that exist in many edge detection algorithms. One can ultimately determine these parameter values automatically by exhaustively testing parameter sets with an objective edge map evaluation measure over a range of images and then generating algorithms based on the analysis of these results.

In general, edge map evaluation measures can be classified as either reference-based or non-reference-based measures. Reference-based edge map evaluation measures require an ideal edge map of an image known as the ground truth in addition to the resultant edge map to make their assessment. Their main advantage is that they can be used to determine edge detector performance within a controlled environment where edge pixel locations are known. Their main disadvantage is that since the ground truth must be known, there are of little use for evaluating edge detector performance on natural images. This is because for natural images, determining the ground truth is non-trivial as ideal edge locations are unknown. Therefore, reference-based edge map evaluation measures are used predominantly for evaluating edge maps of synthetic images with derived ground truths. Moreover, the performance of edge detectors on synthetic images do not necessarily correlate to their performance on natural images, as edge detection kernels are constructed to detect “real-world” edges taking into consideration noise and deviation from edge models. Reference-based objective edge map evaluation measures include probabilistic measures and Pratt’s Figure of Merit [5].

Alternatively, non-reference-based edge map evaluation measures only use information from the resultant edge map and the original image itself to make their evaluation. The obvious advantage of non-reference-based edge map evaluation measures is that no ground truth is necessary. Therefore, they can be used to assess the performance of edge detectors on non-synthetic images. However, in their current stage, non-reference-based edge map evaluation measures suffer from many biases or are restricted in their use. Most are highly unreliable due to the fact that edge detection quality can be very subjective and difficult to quantify. Many of the methods cannot effectively incorporate the original image data. Visual comparisons are highly subjective and cannot be automated [6]. Task-based evaluations [7] quantify edge detector performance based on how well an edge detector output aids a certain task. While this is a useful evaluation, it does not fit under a universal model and depending on the application, they also cannot be automated. Probabilistic measures attempt to estimate a ground truth from multiple edge detector outputs and determine the output which best balances specificity and sensitivity [6]. This method suffers from its bias regarding the generation of the estimated ground truth, as the candidate edge maps used directly affect the estimated ground truth which is determined. Therefore, if the majority of the edge maps used are not of adequate quality or fail to extract certain edge structures which are detected by only a select few of the edge maps, this will be reflected in the derived estimated ground truth. Also, since the original image data is not used, there is no telling how well the best edge detector output determined by this approach corresponds to the original image. Edge connectivity and width uniformity measures assess the quality of edge detector outputs in terms of the formation of proper edge lines [8]. While generating continuous and single pixel width edges are certainly essential criteria for edge detection, edge connectivity and width uniformity measures do not take into account the original image data. Thus, once again, it would be possible to construct a high-scoring edge map independent of actual input data.

Recently, reconstruction-based measures have been developed by the authors [9, 10]. Such measures determine the quality of an edge map based on how well the image can be reconstructed using the original data only at the location of edge pixels in the edge map. This provide an effective means of incorporating the original image data, but it also tends to favor more edge pixels as it allows for more of the original image data to be used in the reconstruction. In order to counter this, a progressive approach was proposed for determining parameter values. However, the progressive approach only provides a means of tuning parameters and not a means of comparing the edge map outputs of various different edge detectors without the edge pixel bias.

In this paper, a new non-reference measure for objective edge map evaluation is introduced. A general formulation for an objective, non-reference edge map evaluation measure is first established. Namely, a measure is defined as the product of a monotonically decreasing function of the number of edge pixels and a function which measures the similarity between the original and reconstructed images (which tends to be a monotonically increasing function of the number edge pixels). A specific instance of the formulation is suggested and used to evaluate the performance of the measure. Two different experiments were performed to illustrate the efficacy of the new measure. In one experiment, the measure is used to determine the best parameters within a single edge detector. In a second experiment, the measure is used to compare the outputs of different edge detectors and determine the best performer. Experimental results over a range of different images illustrate the effectiveness of the presented measure.

The remainder of this paper is organized as follows: Section II provides background information regarding reconstruction-based edge map evaluation methods. Section III describes the new measure, including the general formulation for an objective, non-reference edge map evaluation measure. Section IV illustrates the performance of the new measure. Section V draws conclusions based on these experimental results.

II. Background Information

The main weakness of most non-reference based edge map evaluation measures is their inability to be automated and effectively incorporate the original image data in making their assessment. Reconstruction based measures solve this problem by determining the quality of an edge map based on how well the image can be reconstructed using the original data only at the location of edge pixels in the edge map. This reconstruction approach was actually first suggested by Carlsson [11] as a method of coding images and has since been adapted for the edge map evaluation problem. The quality of the reconstructed image turns out to be is a function of the edge pixel locations and the number of edge pixels in the edge map. Conveniently enough, the quality of the reconstructed image increases as the edge pixel locations in the edge map are closer to their ideal location. This is to say that if one is reconstructing the image using only a portion of the image data, the data at locations of large variations in the image would contain more useful information and yield a better reconstructed image. This forms the main rationale for the use of image reconstruction as a means for edge map evaluation. However, the quality of the reconstructed image also generally increases as the number of edge pixels increase. This is because more information from the original image data can be used to reconstruct the image. Specifically, in the case of thresholding gradient images using incrementally larger thresholds, the quality of the reconstructed image will always degrade as the threshold is increased. This is obviously an unfortunate consequence in terms of the use of image reconstruction for non-reference based edge map evaluation.

A generalized block diagram of the established reconstruction-based measures is shown in Figure 1. For an image I, the reconstruction process starts by first dilating the output edge map e, yielding the dilated edge map e_D. The original image data at the pixel locations $(i,j) \in e_D$ is referred to as the edge tube t, and is used as interpolating values to reconstruct the image. Practically, the edge tube t can be determined by simply multiplying the dilated edge map e_D by the original image I. The interpolation is carried out until all pixels contain an interpolated value. The reconstructed image is then compared to the original image using a similarity measure, which in turn is an assessment of the edge map e. Thus, under this paradigm, the quality of the measure is improved by improving the quality of the reconstruction estimation algorithm and the quality of the similarity measure used to compare the original image and the reconstructed image.

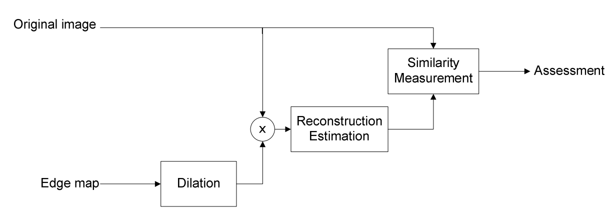


Figure 1 – Block diagram of reconstruction-based measures

A. Reconstruction algorithms

In the reconstruction algorithm suggested by Carlsson, the reconstructed image r minimizes the discretized version of the functional

$$\iint \left(\frac{\partial r}{\partial i}\right)^2 + \left(\frac{\partial r}{\partial j}\right)^2 didj \tag{1}$$

under the constraints

$$r(i,j) = t(i,j) \qquad e_D(i,j) = 1$$

A rudimentary implementation of this is carried out using linear interpolation and maintains the ability to numerically compare relative output. For each pixel location $(i,j) \notin e_D$, the algorithm searches in the four cardinal directions and four intermediate directions for the nearest pixel in the given direction that $\in e_D$. The inverses of the distances of the first pixel encountered in each direction from the given pixel d_k are then used as weights for the weighted average of their respective image intensity values t_k, yielding the reconstructed intensity value for the given pixel. Thus, reconstruction is carried out for each pixel location $(i,j) \notin e_D$ by

$$r(i,j) = \frac{\sum_{k=1}^{8} \frac{t_k}{d_k}}{\sum_{k=1}^{8} \frac{1}{d_k}} \tag{2}$$

The described linear interpolation algorithm can be implemented by its median-based analogue. Namely, the weighted average calculation can be replaced by a weighted median calculation. In doing so, the reconstructed pixels only take on intensity values from found in the edge tube. The weighted median of a sequence x with weights w is given by

$$median_W = median\left(\{x_1 \Diamond w_1, x_2 \Diamond w_2, \ldots, x_{k-1} \Diamond w_{k-1}, x_k \Diamond w_k\}\right) \tag{3}$$

where $\Diamond$ is the replication operator defined as

$$x \Diamond w = \begin{cases} \{x, x, x, \ldots\} & (w \quad times) \quad w > 0 \\ \{-x, -x, -x, \ldots\} & (-w \quad times) \quad w < 0 \end{cases} \tag{4}$$

In order to yield integer weights for reconstruction, the weights w_k are given by

$$w_k = round\left(\frac{100}{d_k}\right) \tag{5}$$

and reconstruction is carried out for each pixel location $(i,j) \notin e_D$ by

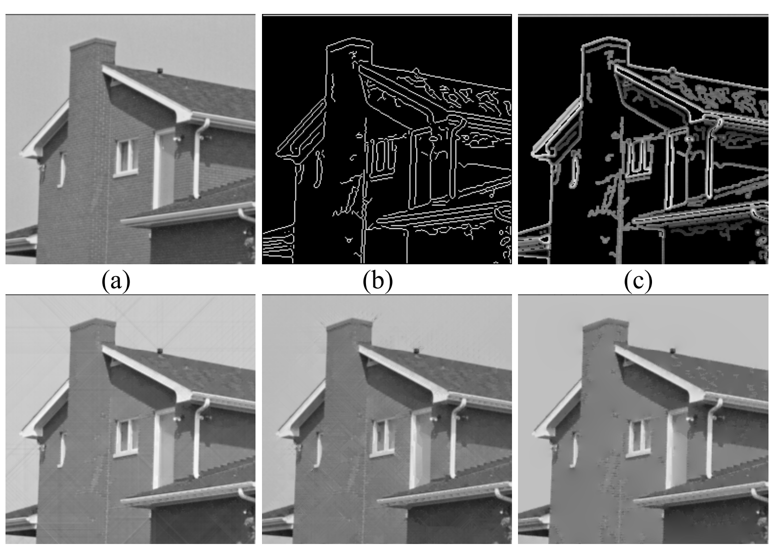
$$r(i,j) = median_W\left(\{t_1 \Diamond w_1, t_2 \Diamond w_2, \ldots, t_7 \Diamond w_7, t_8 \Diamond w_8\}\right) \tag{6}$$

Lastly, many inpainting algorithms have been proposed by Bertalmio et al. [12] based on partial differential equation (PDE) discretizations. In such methods, high order PDEs are designed to restore smooth regions as well as thin structures [12, 13]. Figure 2 shows reconstruction results using the interpolation which have been described. Note the improved performance yielded by the REMI algorithm. It circumvents the creation of false edges during reconstruction, retains the sharpness of the edge tube, is robust to noise, and produces visually more pleasing results [10].

B. Similarity measures

A simple metric to measure the difference between two signals is the mean squared error. The lower the mean squared error between two images, the more similar the two images are. The mean squared error between two images x and y is calculated by

$$MSE(x,y) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n}\left(x(i,j) - y(i,j)\right)^2}{mn} \tag{7}$$



(a) (b) (c) (d) (e) (f)

Figure 2 – (a) Original "house" image, (b) edge detection result using the Canny edge detector, (c) edge tube, reconstruction results using (d) linear interpolation, (e) REMI, (f) and PDE-based inpainting

Though the MSE has a clear and simple to understand mathematical interpretation, it cannot be used to reliably measure the similarity between two images as it does not model the human visual system (HVS) [14]. In general, the MSE lacks the ability to assess image similarity across distortion types.

Bovik's SSIM index [14] defines the similarity of two images as a function of luminance, contrast, and structure. Given two images x and y, it is defined as

$$SSIM(x,y) = \left(\frac{2\mu_x\mu_y}{\mu_x{}^2+\mu_y{}^2}\right)\cdot\left(\frac{2\sigma_{xy}}{\sigma_x{}^2+\sigma_y{}^2}\right) \tag{8}$$

where μ_x and μ_y are the sample means of x and y, respectively, σ_x and σ_y are the sample standard deviations of x and y, respectively, and σ_{xy} is the sample covariance of x and y. Generally, the SSIM is calculated on non-overlapping local windows of the two images to be compared. The mean of these SSIM values (MSSIM) over the entire image is then used to measure the similarity between two images. For two images X and Y, it is defined as

$$MSSIM(X,Y) = \frac{1}{N}\sum_{j=1}^{N} SSIM(x_j, y_j) \tag{9}$$

where M is the number of windows used, and x_j and y_j are the j^{th} window of X and Y respectively. Lastly, in order to improve the dynamic range of the measure for the edge map evaluation problem, an improved SSIM has been suggested [10], given by

$$IMSSIM(X,Y) = 20\cdot MSSIM(X,Y)^{10} \tag{10}$$

C. *Progressive approach*

The reconstruction-based approach provides an effective means of incorporating the original image data, but it also tends to favor more edge pixels as it allows for more of the original image data to be used in the reconstruction. In order to counter this, a progressive approach was proposed based on the difference in the measure values as edge detection parameter values were incrementally increased [12]. The rationale for this is that near optimal parameters, the quality of the reconstruction should not change substantially. Therefore, the difference in the measure values can be used to find such parameters. However, the progressive approach only provides a means of tuning parameters and not a means of comparing the edge map outputs of various different edge detectors without the edge pixel bias.

III. A Generalization For Non-Reference Objective Edge Map Evaluation And New Measure

The non-reference reconstruction-based measures have the clear advantage that they effectively incorporate original image data and can be used to assess the performance of edge detection outputs of non-synthetic images. One would like to use such a measure to make comparisons of results using the same edge detector with different parameters as well as to make comparisons of results using different edge detectors, while also avoiding the unwanted effect of edge maps with more edge pixels yielding higher measure values in all cases. In order to do so, a general model for reconstruction based measures is established. In general, for an input image I with edge map e reconstructed with reconstruction method Φ, a reconstruction-based edge measure can be defined as the exponentially weighted product of a similarity comparison function $f_{S,\Phi}(I,e)$ and an edge pixel density function $f_N(e)$. This is given as

$$measure(I,e) = \left[f_{S,\Phi}(I,e)\right]^{\alpha}\cdot\left[f_N(e)\right]^{\beta} \tag{11}$$

where α and β are exponential weights. The similarity comparison $f_{S,\Phi}(I,e)$ is a function which describes the similarity between the original image and the image which has been reconstructed using method Φ. It is generally a monotonic function with respect to the number of edge pixels in the edge map. The edge pixel density function $f_N(e)$ is a monotonic function of the number of edge pixels in the edge map. If $f_{S,\Phi}(I,e)$ is a monotonically increasing function with respect to the number of edge pixels in the map, then $f_N(e)$ should be chosen to be a monotonically decreasing function of the number of edge pixels in the edge map. Furthermore, it is imposed that both of the functions range from 0 and 1 in order for them to be able to be properly weighted exponentially and for the measure itself to range between 0 and 1. This also eliminates the need for one of the weights, as the ratio of β/α becomes the only weight of practical importance. As a result, α is always set to 1. Effectively, the measure assesses edge maps by how many edge pixels are needed in order to produce a certain degree of structural similarity between the original image and the reconstructed result, with α serving as a penalty parameter for detecting too many edge pixels. This is to say that ideally, the best edge map result consists of the least number of edge pixels at their correct locations needed to characterize all of the relevant structures in the image.

The pair of functions $f_{S,\Phi}(e)$ and $f_N(e)$ used to demonstrate this generalized model are given respectively as

$$f_{S,\Phi}(e) = MSSIM(I,\Phi(I,e)) \tag{12}$$

$$f_N(e) = \left(\frac{mn-N}{mn}\right)^2 \tag{13}$$

where N is the number of edge pixels in the edge map e, m and n are the dimensions of I (and consequently, the dimensions of e), and where the REMI algorithm is used to reconstruct the image from its edge map. The measure then becomes

$$measure(I,e) = MSSIM(I,e)\cdot\left(\frac{mn-N_0}{mn}\right)^2 \tag{14}$$

where once again, the REMI algorithm is used to reconstruct the image I from its edge map e. In this case, the higher the measure value, the better the edge map. For many edge map outputs, the best performer is selected by determining which output has the highest corresponding measure value. Unlike previously established methods, the same procedure can be used to determine optimal parameter values within a single edge detector as well as to determine the best edge detector for a given image. This is because the assessment is truly only a function of the input image I and the edge map e to be assessed.

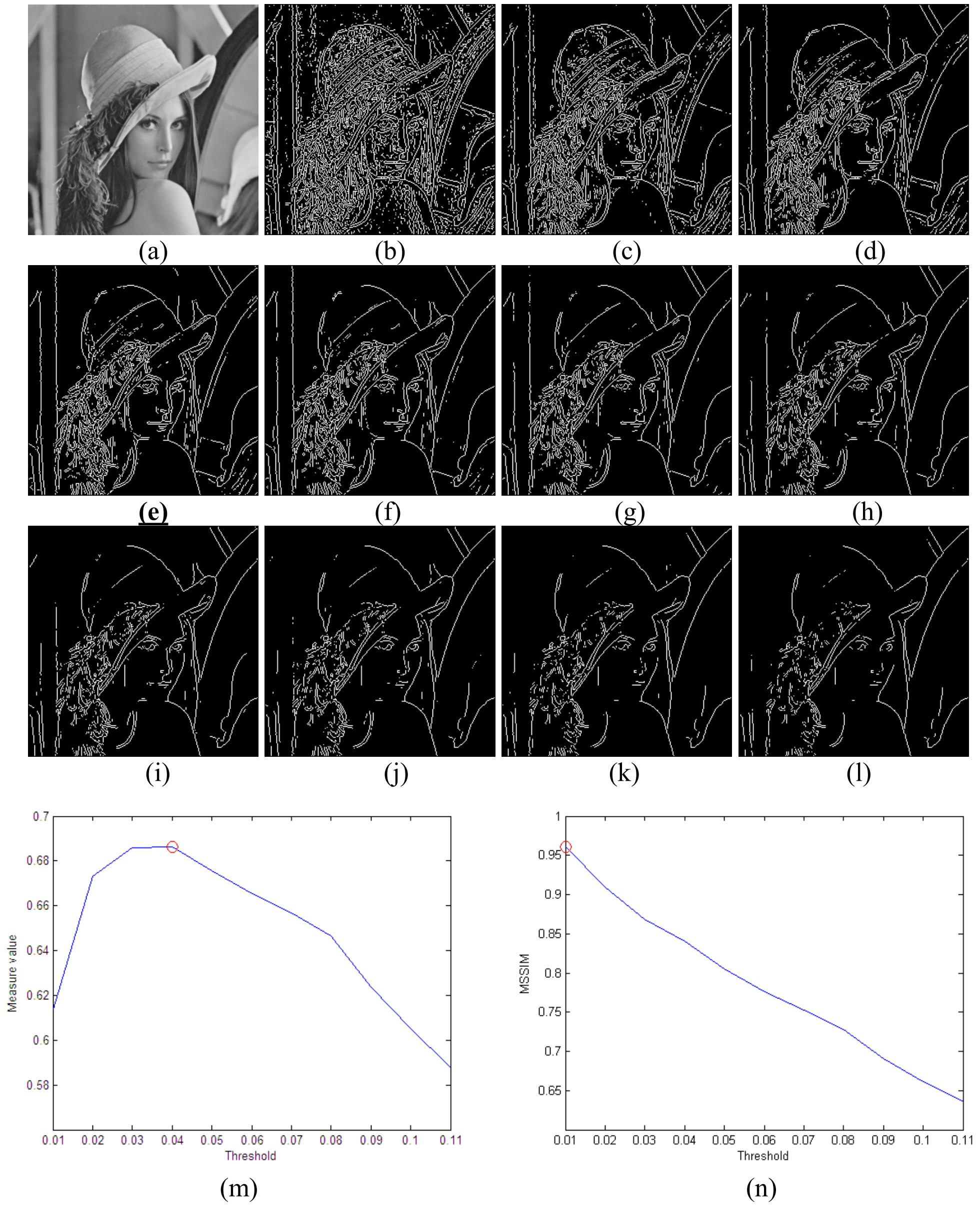


Figure 3 – (a) Original "Lena" image, (b)-(l) edge detection results using the Sobel edge detector with the threshold T ranging from .01 to .11, $\Delta T = .01$ and the top performer according to the presented measure highlighted, (m) presented measure plot indicating $T = .04$ as optimal parameter value, (n) previously established reconstruction-based measure plot using only MSSIM

IV. Experimental Results

The quality of the new measure is assessed by determining how closely it corresponds to human evaluation. This is demonstrated in the form of two experiments. In the first experiment, edge detection is performed using the Sobel edge detector with the thresholding parameter T being incrementally increased, and the best value of T for a given image is determined by the new measure. Specifically, the threshold T ranges from .01 to .11 with $\Delta T = .01$. In a second experiment, edge detection is performed using the Roberts, Prewitt, Sobel, Frei-Chen, Canny, and LoG edge detection algorithms, and the best edge detector output for a given image is determined using the new measure. These results are then compared to human evaluation. Both tests were performed on a variety of different images of different image classes with varying size and amounts of complexity in the scene.

An example of the use of the measure as a means of determining optimal parameter values is shown in Figure 3. Figure 3(n) illustrates the need for the edge pixel density function as the use of the MSSIM alone results in the discussed edge pixel bias. Similarly, an example using the measure as a means of determining the best edge detector output is shown in Figure 4. Table 1 summarizes a portion of these experimental results comparing the presented measure to human evaluation. Often times, the best parameter value determined by human evaluation is not clear. One could argue that two different values appear to balance extracting details and eliminating spurious edge responses in the output edge map. In such cases, the measure value between the two debated parameter values

Image class	Image	Best Sobel thresholding parameter value *T* chosen by		Best edge detector chosen by	
		Human Evaluation	Presented Measure	Human Evaluation	Presented Measure
Miscellaneous	Lena	.03, .04	.04	Canny	Canny
	Cameraman	.04, .05	.05	Canny	Canny
	Clock	.04	.04	Canny	Canny
Medical	Brain MRI	.04, .05	.05	Canny	Canny
	Abdomen MRI	.04	.04	Sobel	Sobel
Aerial	Aerial	.09	.09	Canny	Canny
	Pentagon	.04, .05	.05	Sobel	Sobel

Table 1 – Summary of experimental results using the presented measure

are usually very close to one another. In all cases, the presented measure selected a parameter value which agreed with human evaluation. Similarly, the best edge detector chosen by the measure also coincides with human evaluation. In most cases, the Canny edge detector clearly outperforms the other methods. However, there are a few instances when the Sobel edge detector extracts more necessary fine detail and outperforms the Canny edge detector.

V. CONCLUSIONS

A new non-reference-based edge map evaluation measure has been presented based on a generalization of reconstruction-based edge map evaluation. The presented measure retains the advantages of reconstruction-based edge map evaluations while eliminating some of the biases of previously established methods. The effectiveness of the new measure has been shown both as a means of comparing the performance of many edge detectors as well as a means of selecting parameters within a single edge detector. Experimental results show that the measure coincides with subjective evaluation, validating the usefulness of the measure.

REFERENCES

[1] J. Parker, Algorithms for Image Processing and Computer Vision, John Wiley & Sons, New York, 1996.

[2] M. Bennamoun and B. Boashash, “A Structural-Description-Based Vision System for Automatic Object Recognition,” IEEE Transactions on Systems, Man, and Cybernetics – Part B: Cybernetics, Vol. 27, No. 6, December 1997.

[3] P.L. Rosin, “Robust Pose Estimation,” IEEE Transactions on Systems, Man, and Cybernetics – Part B: Cybernetics, Vol. 29, April 1999. D. Nguyen, D. Halupka, P. Aarabi, and A. Sheikholeslami, “Real-Time Face Detection and Lip Feature Extraction Using Field-Programmable Gate Arrays,” IEEE Transactions on Systems, Man, and Cybernetics – Part B: Cybernetics, Vol. 36, No. 4, August 2006.

[4] W. Pratt, Digital Image Processing, John Wiley & Sons, New York, 2nd edition, 1991.

[5] Y. Yitzhaky and E. Peli, “A Method for Objective Edge Detection Evaluation and Detector Parameter Selection,” IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 25, No. 10, October 2003.

[6] M. Shin, D. Goldgof, K Bowyer, and S. Nikiforou, “Comparison of Edge Detection Algorithms Using a Structure from Motion Task,” IEEE Transactions on Systems, Man, and Cybernetics – Part B: Cybernetics, Vol. 31, No. 4, August 2001.

[7] Q. Zhu, “Efficient evaluations of edge connectivity and width uniformity,” Image Vision Computing, Vol. 14, p. 21-34, February 1996.

[8] B. Govindarajan, K. Panetta, and S.Agaian, "Progressive Edge Detectionon Multi-bit Images Using Polynomial-based Binarizations," IEEE Int. Conf. Mach. Learning and Cyber., July 2008.

[9] B. Govindarajan, K. Panetta, and S.Agaian, "Image Reconstruction for Quality Assesment of Edge Detectors," Sub.: IEEE Int. Conf. on Systems, Man, and Cybernetics, October 2008..

[10] S. Carlsson, “Sketch based coding of gray level images,” *IEEE Trans. Image Processing*, Vol. 15, No. 1, p. 57-83, Jan. 1988.

[11] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester, “Image inpainting,” in *Proc. ACM SIGGRAPH 2000*, pp. 417-424.

[12] C. Ballester, M. Bertalmio, V. Caselles, G. Sapiro, and J. Verdera, “Filling-in by joint interpolation of vector fields and gray levels,” *IEEE Trans. Image Processing*, vol. 10, no. 8, pp. 1200-1211.

[13] Z. Wang, E. P. Simoncelli and A.C. Bovik, “Multi-scale structural similarity for image quality assessment,” IEEE Asilomar Conf. Signals, Systems, and Computers, 2003.