A Quantitative Methodology for Analyzing the Performance of Detection Algorithms

†T. Kanungo, †M. Y. Jaisimha, ‡J. Palmer and †R. M. Haralick
{tapas,jaisimha,haralick}@ee.washington.edu, jpalmer@u.washington.edu
†Department of Electrical Engineering, FT-10
‡Department of Psychology, NI-25
University of Washington
Seattle, WA 98195, USA

Abstract

There has been increasing interest in quantitative performance evaluation of computer vision algorithms. The usual method is to vary parameters of the input images or parameters of the algorithms and then construct operating curves that relate the probability of mis-detection and false alarm for each parameter setting. Such an analysis does not integrate the performance of the numerous operating curves. In this paper we outline a methodology for summarizing many operating curves into a few performance curves. This methodology is adapted from the human psychophysics literature and is general to any detection algorithm. The central concept is to measure the effect of variables in terms of the equivalent effect of a critical signal variable. We demonstrated the methodology by comparing the performance of two line detection algorithms.

1 Introduction

There has been increasing interest in quantitative performance evaluation of computer vision algorithms. This is especially important in order to compare the performance of dissimilar algorithms on a common quantitative basis. The usual method is to vary parameters of the input images or parameters of the algorithms and then construct operating curves that relate the probability of mis-detection and false alarm for each parameter setting. Such an analysis does not integrate the performance of the numerous operating curves. In this paper we outline a methodology for summarising many operating curves into a few performance curves. This methodology is adapted from the human psychophysics literature and is general to any detection algorithm.

Performance analysis is difficult. The question that arises immediately is – How exactly does one define performance? Issues that need to be addressed are: (i) What image population is relevant? (ii) Is the performance evaluated independent of the algorithm? (iii) How are differences in performance measured? An area with previous work on quantitative performance evaluation is in edge detection and thresholding. [5, 1, 18, 16, 21, 14, 11, 20]. Most of the papers present an analysis that is specific to edge detection. Furthermore, the performance is finally a number, e.g., percentage of edge points detected, etc. There is little further analysis of the sensitivity of performance to relevant factors such as the context of the edge.

In this paper, we present a methodology for designing experiments to characterize detection algorithms. We adopt an established methodology that has been used and tested in psychophysics. The central concept is to measure the effect of variables in terms of the equivalent effect of a critical signal variable. For example, psychophysicists study the performance of humans in the task of edge and grating detection by measuring the contrast necessary for detection under a variety of conditions [3, 8, 7]. The effect of the various conditions is measured by the equivalent effect of contrast as quantified by the contrast threshold. A more detailed report of the work presented here can be found in [15].

Section 2 describes the general performance evaluation methodology. The example experiment we perform to demonstrate the methodology is described in section 3. Here we discuss the detection tasks, two algorithms for detecting our targets, and describe the population of images the algorithms and the experimental protocol. Section 4 summarizes all the results. The benefits of our methodology and its application to other detection problems is discussed in section 5.

2 Data Analysis Methodology

In a typical detection task, the system is required to report the presence or absence of a target in an input image. In any detection task there are some variables that affect the signal to noise ratio S/N, in the image. For example, edge contrast in edge detection. The effects of all other variables are measured in terms of the S/N variable. The methodology is applied in four steps. The first two are standard decision analysis [1], the last two are inspired by psychophysical methods [7].

Step 1: A large number of input images, both with and without target, are created under fixed conditions. They differ only by the effects of noise. Each image is then provided as input to the system whose performance is to be evaluated. The output of the system is a number which is a measure of the evidence of the presence of a target. This number is referred to as the *evidence strength*. Each evidence strength has an associated frequency (i.e. the number of times it appears at the output of the system in the course of the experiment). The frequency count is plotted versus the evidence strength. Two frequency distributions are obtained, one for the case of target present and one for the case of no target present.

Step 2: In order to decide whether or not the system has indicated the absence or presence of a target, an evidence criterion C has to be applied to the evidence strengths output by the system. In order to study the performance of the system, the evidence criterion is varied through a set of values. For each value of the evidence criterion there are corresponding values of probability of misdetection given a a target, $P(M|target)$, and probability of false alarm given no target, $P(F|no-target)$. The plot of $P(M|target)$ versus $P(F|no-target)$ as the evidence criterion is varied is called the *operating characteristic*. For the equal bias case, choose the operating criterion, C_0 as the evidence strength for which $P(M|target) = P(F|no-target)$. The equal bias probability of error, $P(E)$, is then defined as $P(E) = (P(M|target) + P(F|no-target))/2$ for equal probability of target and no-target.

Step 3: For different values of the signal to noise ratio, repeat steps 1 and 2. Each value of S/N results in an operating characteristic from which the equal cost probability of error, $P(E)$, can be determined. Plot $P(E)$ versus S/N and choose C_T, the value of S/N for which $P(E) = 0.25$, as the *contrast threshold*. The value of $P(E)$ corresponding to the C_T is chosen half way between pure chance ($P(E) = 0.5$) and perfect ($P(E) = 0.0$). If the particular application so demands, the value of $P(E)$ that defines the contrast threshold C_T can be chosen differently.

Step 4: For different values of the variable of interest V, repeat steps 1, 2 and 3 and plot contrast threshold C_T versus the variable of interest V. This curve now characterizes the effect of the variable of interest in terms of the effect of contrast. The effect of any variable can be measured by its effect on the contrast threshold.

In summary, steps 1 and 2 result in an unbiased measure of performance using standard decision analysis. Step 3 provides a measure in terms of a signal variable at a fixed S/N ratio. Step 4 measures the effect of any other variable by the common currency of the signal variable.

3 The Experiment

3.1 The Detection Task

We illustrate the general methodology with an analysis of a particular issue in edge and line detection. The question raised is – How selective is an edge detection algorithm to irrelevant edges? For example, can the algorithm detect an edge in the context of a texture full of oriented edges? A similar problem has been analyzed for human detection performance in [3]. The input is either a target image or a non-target image. The target image consists of a vertical edge at a known column position, irrelevant square wave grating at various orientations, and Gaussian noise. The non-target image consists of only the grating and the Gaussian noise – no vertical edge. The task is to detect the presence or absence of this edge and ignore the grating.

3.2 The Algorithms

3.2.1 Algorithm 1

This algorithm also has two stages: an edge detection stage followed by a line detection stage. For edge detection we use the second directional derivative facet edge detector [9]. For the second stage, the line detection is done by the Burns line finder [2].

The Burns line finder takes the output of the edge detector and labels all the connected lines it finds. Next, we count the number of pixels in the middle three columns that were classified by the Burns line finder as line pixels and use the count as the evidence strength.

3.2.2 Algorithm 2

This algorithm has two stages: an edge detection stage followed by a line detection stage. For edge detection, as in Algorithm 1, we use the second directional derivative facet edge detector [9]. The line detection is performed using the Hough transform technique [13].

The line detection stage takes the output of the edge detector and maps the edge pixels to the distance-angle Hough space. Since, we know in advance the exact location of the edge, only the pixels in the vicinity of the edge need to vote for the edge.

3.3 Experimental Protocol

The images used were 513×513 pixels. The grating has a 50% duty cycle and has a half period W of 16 pixels. The phase ϕ of the grating is the offset of the rising edge of the grating from the rising edge of the vertical edge of interest. When the offset is zero, the phase is 0 degrees and the grating has a constructive interference. When the offset is equal to W, the phase is 180 degrees, and the grating has a destructive interference on the vertical edge. For the experiment the phase is fixed at 180 degrees, i.e., the offset was W. The orientation θ of the grating with respect to the edge is varied through a set of possible angles. The rotation of the grating is around the center pixel of the image. The contrast C_e of the edge and the grating C_g (the step size in terms of fractions of the mean gray value L_0) are also varied. Similarly the standard deviation, σ_η, of the Gaussian noise also takes up values that are fractions of L_0. The mean, μ_η, of the noise is assumed to be zero. The values of the parameters of the experiment are as follows: $L_0 = 100$, $C_m = 10$ % of L_0, $W = 16$ pixels, $\sigma_\eta = 20\%$ of L_0, $\mu_\eta = 0$, $\phi = 180$ degrees (destructive phase), $C_e = 2\%$ of $L_0, \ldots$, 26% of L_0, $\theta = 0, 1, 3, 5, 45, 90$ degrees. Two sample images used in the experiment are shown in Figure 1.

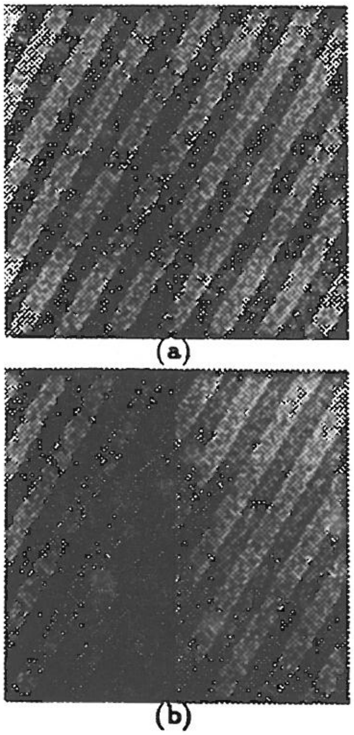

Figure 1: Sample images for the experiment. (a) This image has only the grating at an orientation of 45 degrees. (b) This image has the grating at the same orientation as (a) as well as the vertical edge.

4 Results

The performance of the line detection algorithm is analyzed using the four steps outlined above.

The first step is to measure the frequency distribution of the evidence strengths for a vertical edge given images with or without an edge. For Algorithm 1, the evidence strength is the count of vertical edge pixels detected by the Burns line finder in the center of the image. The frequency histograms are shown for the case with $\sigma_\eta = 20$, $\theta = 45$ degrees in Figure 2. As expected, the edge images result in a distribution with higher evidence strength values and this appears to the right of the graph.

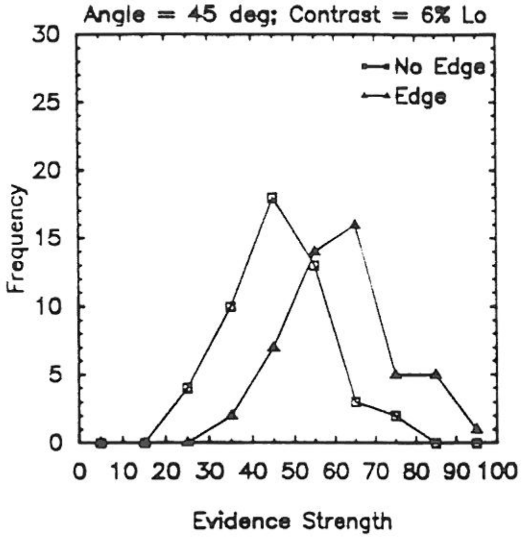

Figure 2: Step 1 – Histogram for evidence strengths for images with and without the vertical edge. In this case the grating angle was 45 degrees, the edge contrast was 6%.

The second step is to measure the operating characteristic from these frequency distributions. Figure 3 shows the probability of false alarms as a function of the probability of misdetection for the full range of possible evidence criteria. As the value of the edge detection criterion is lowered, the probability of false alarm decreases but the probability of misdetection increases. We use a non-negative least square optimization technique to fit monotonically decreasing smooth curves [12]. The operating criterion is chosen as the point for which $P(M|edge) = P(F|no-edge)$. The probability of error $P(E) = (P(M|edge) + P(F|no-edge))/2$. For example, the probability of error is 0.29, when $C_e = 6\%L_0$ in 3.

The third step is to measure the probability of error as a function of the signal to noise ratio. For this experiment, the signal to noise ratio was manipulated by varying the vertical edge contrast, C_e. The results for many orientations, θ, are shown in Figure 4. This curve falls from a maximum expected error of 0.5 to no errors as the edge contrast increases. If the distributions found with step 1 are equal variance Gaussian, these functions will be cumulative Gaussians. Such a function is roughly linear in its middle range and this approximation is used to

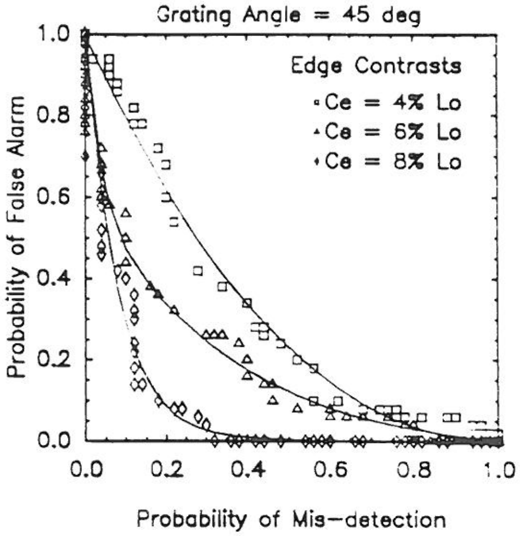

Figure 3: Step 2 – Operating curves for grating angle of 45 degrees and edge contrasts of 4%, 6%, and 8%.

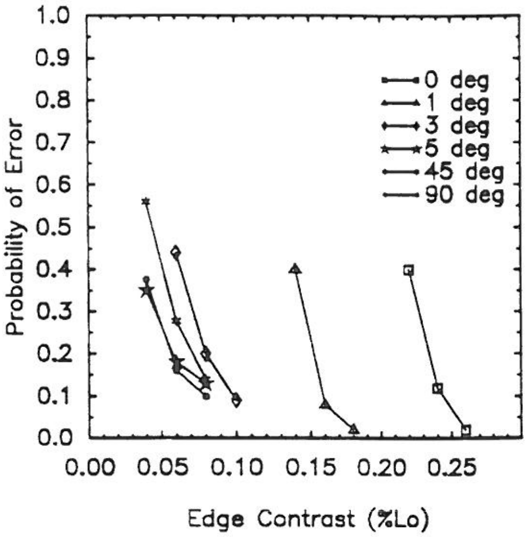

Figure 4: Step 3 – The effect of contrast on the equal bias probability of error. The contrast threshold, C_T, is the contrast required for a 25% error rate.

extract the contrast threshold, C_T, at performance equal to 0.25 error. Contrast threshold C_T is the edge contrast required to get a 25% error rate or equivalently 75% detection rate. More precise results can be based on [6]. From this threshold and a template of these functions, one can estimate the probability of error for any contrast.

The fourth step is to measure the effect of the orientation of the irrelevant grating. Figure 5 shows the contrast threshold for several orientations. It can be seen that for grating orientations greater than 5 degrees and less than 90 degrees, you need a vertical edge contrast of 6% of L_O to have a 75% detection rate. The performance is remains relatively constant over this range of grating orientations. There is a sudden deterioration of performance as the grating orientations becomes smaller that 5 degrees. In fact, when the grating orientation is 0 degrees, i.e. the grating is in destructive phase, an edge contrast of 23% is required in order to get a 75% detection rate.

The same analysis was done for Algorithm 2. The results are plotted along with the results of Algorithm 1 in Figure 5. We can see that both algorithms have similar worst case performance (grating orientation of 0 degrees) – both need approximately 23 % contrast for 75% detection rate. But, Algorithm 1 has a better asymptotic performance. That is, when the orientation of the grating is greater than 5 degrees, Algorithm 1 needs about 6% edge contrast whereas Algorithm 2 needs about 13% edge contrast. Finally, the performance of both algorithms start deteriorating at around 5 degrees.

Without doing the analysis outlined above, it would have been difficult to hypothesize the worst and asymptotic performance of the two algorithms from few operating curves of step 2. In addition, it would have been difficult to predict the location of the knee of the curve. We are currently developing statistical measures for each step in this analysis [10, 6].

5 Discussion

5.1 Advantages of using thresholds

This methodology follows others in using decision analysis to combine the two kinds of errors into a single error probability given an decision criterion. The current analysis extends this by manipulating a signal-to-noise variable to measure a threshold as is common in psychophysics. Thresholds have several advantages as performance measures:

- Thresholds are defined independent of the algorithm. In our case we used a contrast threshold that gave us a 75% detection rate.
- By defining threshold at a fixed performance rate, one can compare the effect of other variables at a known and comparable S/N ratio.
- Thresholds vary over as large a range as the signal-to-noise ratio. Thus, thresholds will have a dynamic range that is not as restricted as the probability of error measure. This is a problem because very low error probabilities cannot be measured in a practical experiment.
- Thresholds can be measured without factorial experiments by adaptively choosing signal-to-noise ratios appropriate for each set of conditions.. For example, if we need the threshold for which the $P(E)$ is 0.25, there is no need to run the experiment with parameters that give S/N for which the $P(E)$ is far from 0.25. This can substantially reduce the size of experiments.

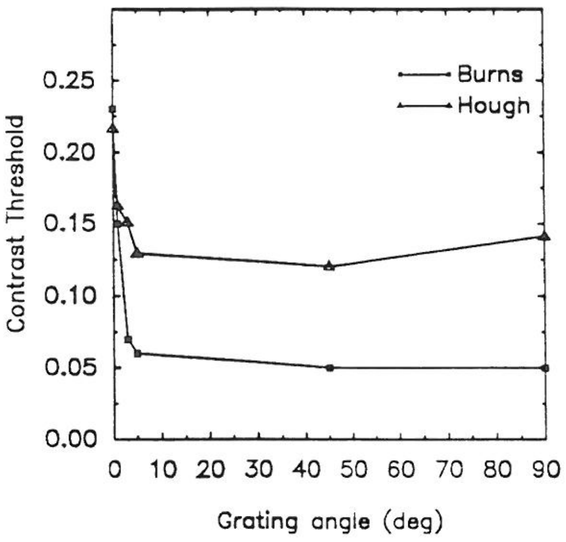

Figure 5: Step 4 – Contrast threshold, the contrast necessary for a 75% detection rate, as a function of grating angle for the two algorithms Facet+Burns and Facet+Hough.

- Thresholds provide a way to measure differences in performance over large ranges. For example, is one algorithm worse than another by a 10% difference in the threshold, a 100% difference, or a 1000% difference? Thresholds give a way of distinguishing between large and small effects.

The use of thresholds is not new in signal processing. The most common example is the the notion of bandwidth. The bandwidth is usually defined as the frequency range within which amplitude response of a filter is is greater than 3db. In this case, the threshold used is the 3db amplitude response.

5.2 Analyzing algorithm behavior and design of better algorithms

Since different algorithms can be compared using this methodology, it can be used a a tool for understanding the behavior of different algorithms. For example, why should one algorithm have better asymptotic performance than other? What parameters in the algorithm control the location of the "knee" of the curve? Can they be modified to suit our requirements? Furthermore, these performance curves can be used to design better algorithms using the good features of various algorithms.

5.3 Summarizing performance curves

The methodology allows one to summarize many operating curve by a few performance curves. In our experiment, each histogram and corresponding operating curve resulted from 100 trials. Each curve in step 3 of the methodology represents 3 operating curves or 300 trials. The final curve in step 4 of the methodology represents $3 \times 6 = 18$ operating curves, or $3 \times\ 6 \times 100 = 1800$ trials. In contrast, most methodologies existing in the literature today provide only the operating curves (step 2 of our methodology). Thus our methodology allows the researcher to convey more information in a meaningful way.

5.4 Analytic performance evaluation

In case an analytic model is available, it is not necessary to run the experiments to compute the operating curve. In fact, the probabilities can be computed form the analytic expression for the probabilities of mis-detection and false alarm [19]. But we still encounter the problem of summarizing the numerous operating curves. Our methodology can be applied, without modification to the analytic results, just as it is applied to the empirical results. Thus the analytic results can be summarized just as we summarized the empirical results. Furthermore, there are cases when either the analytic model of an algorithm is not numerically tractable or is not known. In such cases it is possible to approach the performance evaluation problem in a quantitative fashion.

5.5 Applications

A strength of the methodology is that it can be applied to any detection problem. The line detection example developed in this paper was for demonstrating the application of this methodology. The key steps to applying this methodology to any algorithm are (i) converting the algorithm into a detection algorithm, and (ii) choosing the appropriate signal variable to use as the threshold.

Another appropriate example is the detection of corners and junctions [22]. To analyze corner, consider an image that does or does-not contain a corner. The corner detection algorithm outputs an evidence strength that indicates whether there is a corner in an image. We can manipulate the angle of the corner to find the signal-to-noise threshold. This performance measure can be used to study the effect of other variables such as the length of the lines making up the corner.

To analyze automatic target detection algorithms, input images would either contain image of the target or no target. The algorithm would first have to detect the presence or absence of the target in the image. Now, one could fix all the variables (distance, shape of target, etc.) except the contrast of the signal. The contrast could be used to control the signal to noise ratio and the variable of interest could be the specular reflectance of the target. One could study how the performance deteriorates as the specular reflectance increases.

In the case of inspection of machined parts, the vision algorithm algorithm decides whether a machine part is satisfactory ("within spec") or not ("out of spec"). This is a detection task. The errors are either misdetection errors or false alarm errors. In this case, the degree of defect in the machined part could be used as the signal variable. For more details see [17].

The task of pose error estimation can be converted into a detection task by asking the question – is the estimated pose of the object within specific error bounds or not. A measure of error could be computed as follows. After the pose of an object is estimated, the average distance between the vertices of the original object and the back-projected object could be used as an error measure. A threshold on this error makes converts the problem into a detection task. For more details please refer to [4].

6 Conclusion

We describe a methodology for characterizing and summarizing the performance of any detection algorithm. It extends the previous applications of decision analysis by the addition of threshold measures inspired from psychophysics. This is a general methodology that can be applied to any detection algorithm.

Acknowledgements
Authors would like to thank Horace Barlow, Harrison Barrett, and Charles Metz for their comments.

References

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

[2] J.B. Burns, A.R. Hanson, and E.M. Riseman. Extracting straight lines. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 8:425–455, July 1986.

[3] F.W. Campbell and J.J. Kulikowski. Orientation selectivity of the human visual system. *Journal of Physiology*, 187:437–445, 1966.

[4] O.I. Camps, L.G. Shapiro, and R.M. Haralick. Recognition using prediction and probabilistic matching. In *Proc. of IEEE/RSJ Fifth International Conference on Intelligent Robots*, Raleigh, North Carolina, July 1992.

[5] E. S. Deutsch and J. R. Fram. A quantitative study of the orientational bias of some edge detector schemes. *IEEE Transactions on Computers*, March 1978.

[6] D.J. Finney. *Probit Analysis*. Cambridge University Press, 1971.

[7] G.A. Gescheider. *Psychophysics: Methods, Theory, and Application*. Erlbaum, Hillsdale, NJ, 1985.

[8] D.M. Green and J.A. Swets. *Signal Detection Theory and Phsychophysics*. Robert E. Krieger Publishing Company, New York, second edition, 1974.

[9] R.M. Haralick. Digital step edges from zero crossings of second directional derivatives. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, January 1984.

[10] R.M. Haralick. Performance assessment of near perfect machines. *Journal of machine vision and applications*, 2:1–16, 1989.

[11] R.M. Haralick and J. S. J. Lee. Context dependent edge detection and evaluation. *Pattern Recognition*, 23(1/2):1–19, 1990.

[12] R.M. Haralick, K.B. Thornton, and T. Kanungo. Monotonic fitting. Unpulished Report. The monotonic curve fitting problem is solved by posing the problem as as constrained non-linear optimization problem, July 1992.

[13] P.V.C. Hough. A method and means for recognizing complex patterns. *U.S. Patent 3,069,654*, 1962.

[14] T. Kanungo, M.Y. Jaisimha, R.M. Haralick, and J. Palmer. An experimental methodology for performance characterization of a a line detection algorithm. In *Proc. of SPIE Symposium on Advances in Intelligent Systems*, Boston, Nov 1990.

[15] T. Kanungo, M.Y. Jaisimha, J. Palmer, and R.M. Haralick. Quantitative performance evaluation of detection algorithms. Internal Report, Feb 1993.

[16] L. Kitchen and A. Rozenfeld. Edge evaluation using local edge coherence. *IEEE Transactions on Systems, Man and Cybernetics*, SMC-11(9):597–605, 1981.

[17] B.R. Modayur, L.G. Shapiro, and R.M. Haralick. Visual inspection of machined parts. In *Computer Vision and Pattern Recognition*, pages 393–398, Champaign, IL, June 1992.

[18] T. Peli and D. Malah. A study of edge detection algorithms. *Computer Graphics and Image Processing*, 20:1–21, 1982.

[19] V. Ramesh and R.M. Haralick. Random perturbation models and performance characterization in computer vision. In *Computer Vision and Pattern Recognition*, pages 521–527, Champaign, IL, June 1992.

[20] P.K. Sahoo. A survey of thresholding techniques. *Computer Vision, Graphics, and Image Processing*, 41, 1988.

[21] Tan, Gelfand, and E.J. Delp. A comparative cost function approach to edge detection. *IEEE Transactions on Systems, Man, Sybernetics*, pages 1337–1349, December 1989.

[22] C.H. Teh and R.T. Chin. On the detection of dominant points on digital curves. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 11:859–872, 1989.