WEIGHTED DCT COEFFICIENT BASED TEXT DETECTION

Su Lu, Kenneth E. Barner

Department of Electrical and Computer Engineering
University of Delaware
Newark, Delaware, 19716

ABSTRACT

This paper addresses text detection utilizing weighted Discrete Cosine Transform (DCT) coefficients as a discrimination statistic. The sum of absolute value of DCT coefficients in each block is considered as an energy statistic for detection. Linear Discriminant Analysis (LDA) is conducted to calculate the optimal threshold. Different types of weights are employed to strengthen discrimination statistic, which include uniform, binary, linear and quadratic weights. The evaluation of weighted algorithms is conducted in the Receiver Operating Characteristics (ROC) space. The ROC curves show that there is a tradeoff between True Positive Rate (TPR) and False Positive Rate (FPR) for all weight configurations. In terms of maximizing the separation between two distributions, experimental results show that the quadratic weighted energy achieves the best recall and precision.

Index Terms— Text detection, DCT, LDA, ROC curve.

1. INTRODUCTION

The automatic retrieval of text information from color images has gained increasing attention in recent years. Text detection can be found in many applications, such as road sign detection, map interpretation, engineering drawings interpretation, etc. [1, 2]. Extensive research efforts have been directed to the detection, segmentation, and recognition of text from still images and video. However, due to the complexity of text appearance in images, text detection is still a difficult and challenging task in image processing.

DCT based methods do not make assumptions about the nature of text in an image, e.g., character alignment, same size of character, etc. As a result, very small font sizes can be detected, even if the text is not easily readable. Crandal [3] performs an 8×8 block-wise DCT on a video frame and extracts a subset of the DCT coefficient from each block. The sum of the absolute values of these coefficients is computed and regarded as a measure of the text energy of that block. The subset of DCT coefficients that best corresponds to the properties of text was determined empirically by trying all combinations of between 1 and 64 coefficients. Instead of choosing an optimal subset empirically, Shiratori [4] selects middle frequency components as the text detection feature. He also employs an edge count filter to remove non-text components. This paper develops a set of weight functions that can strengthen discriminant statistic, thus enhancing the detection performance. The study of the different types of weight functions demonstrates that each weight has its advantage in different sensitive cases. Overall, the quadratic weighted method achieves highest recall and precision rates, which is based on fixed optimal threshold generated by LDA.

The remainder of the paper is organized as follows. In Section 2, we discuss transform domain text detection algorithms in detail, which include motivation for employing DCT, linear discriminant analysis, weighted detection and filtering detection. Evaluation metrics for each proposed algorithms are presented in Section 3. Finally, conclusions are drawn in Section 4.

2. DCT DOMAIN TEXT DETECTION

It is known that DCT has a strong energy compaction property [5, 6], i.e., most of the signal information tends to be concentrated in a few low-frequency components. In image processing, the DCT exhibits excellent energy compaction for highly correlated images. Clearly, an uncorrelated image has its energy spread out, whereas the energy of a correlated image is packed into the low frequency region. For text detection, the text is usually uncorrelated with the image background. Hence, it is possible to make use of this fact to distinguish text blocks from natural image blocks.

In our detection procedure, the image is block-wise DCT transformed and the sum of absolute value DCT coefficients in each block is consider as an energy statistic for detection. Most natural image energies are concentrated in low frequency components while the energies of text are concentrated in high frequency components. Based on this fact, DCT blocks are divided into two classes, text and non-text, by comparing the weighted energy to a threshold. LDA (Fisher's linear discriminant) is conducted to generate an optimal threshold and the blocks whose feature are greater than the threshold are regarded as text blocks. To find such a threshold, we discuss linear Fisher discriminant analysis in next section.

2.1. Fisher's Discriminant Threshold

Fisher's linear discriminant considers not only between-class variation but also within-class variation, and optimizes the solution by maximizing the ratio of between-class scatter to within-class scatter. Fisher defined the separation between two distributions to be the ratio of the variance between the classes to the variance within classes:

$$S = \frac{\sigma^2_{between}}{\sigma^2_{within}} \tag{1}$$

In our case, $n = 2$ and the ratio can be written as

$$S = \frac{\omega_0\omega_1(\mu_0 - \mu_1)^2}{(\omega_0 + \omega_1)(\omega_0\sigma_0^2 + \omega_1\sigma_1^2)} \tag{2}$$

where μ_i and σ_i^2, $i = 0, 1$ are the mean and the variance of the ith class, and $\omega_i, i = 0, 1$ is the number of occurrence of each class. The optimal division $\omega_{opt} = (\omega_0^*, \omega_1^*)$ is achieved when the maximum separation occurs, i.e, the ratio S reaches its maximum.

To obtain corresponding optimal threshold, we sort the energy of each block in ascend order. Then, the threshold is given by

$$T = \frac{E_{(\omega_0^*)} + E_{(\omega_0^*+1)}}{2} \tag{3}$$

where E is block energy and the subscript (ω_0^*) enclosed in parentheses indicates the $\omega_0^* th$ order statistic of E. This threshold is utilized to distinguish the text blocks from non-text blocks.

2.2. Weighted Text Detection

To improve the discrimination of text blocks, we introduce weighting into the energy statistic calculation. Specifically, we consider uniform, binary, linear, and quadratic weights. Assume block size is $N \times N$. Then, we can obtain weighted block energy:

$$E = \sum_{p=0}^{N-1}\sum_{q=0}^{N-1} X(p,q) \times W(p,q) \quad p,q = 0, \ldots, N-1 \tag{4}$$

where $W(p,q)$ is weight matrix element and $X(p,q)$ is the DCT coefficient of each image block.

2.2.1. Uniform Weight

For coefficients of an 8×8 block-wise DCT transform, the uniform weight is given by

$$W_U(p,q) = \begin{cases} 0\,, & p = q = 1 \\ 1\,, & 1 < p + q \leq 16 \end{cases} \tag{5}$$

From this point forward, all weight functions are based on an 8×8 block-wise DCT transform.

2.2.2. Binary Weight

In many cases, text regions only show differences in some DCT frequency regions. This fact makes using a subset of the transform coefficients reasonable, which is equivalent to binary weighting. Shiratori [4] makes use of middle frequency part of the DCT coefficients. This binary weighting, denoted by W_{BM}, is given by

$$W_{BM}(p,q) = \begin{cases} 0\,, & 0 < p + q \leq 7 \\ 1\,, & 7 < p + q \leq 10 \\ 0\,, & 10 < p + q \leq 16 \end{cases} \tag{6}$$

Also of interest are the low and high frequencies. The corresponding binary weight functions, W_{BL} and W_{BH}, are given by

$$W_{BL}(p,q) = \begin{cases} 0\,, & p = q = 1 \\ 1\,, & 1 < p + q \leq 7 \\ 0\,, & 7 < p + q \leq 16 \end{cases} \tag{7}$$

$$W_{BH}(p,q) = \begin{cases} 0\,, & 0 < p + q \leq 10 \\ 1\,, & 10 < p + q \leq 16 \end{cases} \tag{8}$$

2.2.3. Linear Weight

Binary weights can reduce computations, but at the cost of increasing detection errors. Since the text blocks have more energy in high frequency components, linearly increasing weights are selected to amplify the high frequency energy while mitigating the effect of low frequency energy. For 8×8 DCT coefficients, linear weights, denoted by W_L are given by

$$W_L(p,q) = \begin{cases} 0, & p = q = 1 \\ \frac{32}{7}(p + q - 2), & p + q > 2 \end{cases} \tag{9}$$

Thus, the elements of W_L increase linearly along the diagonal.

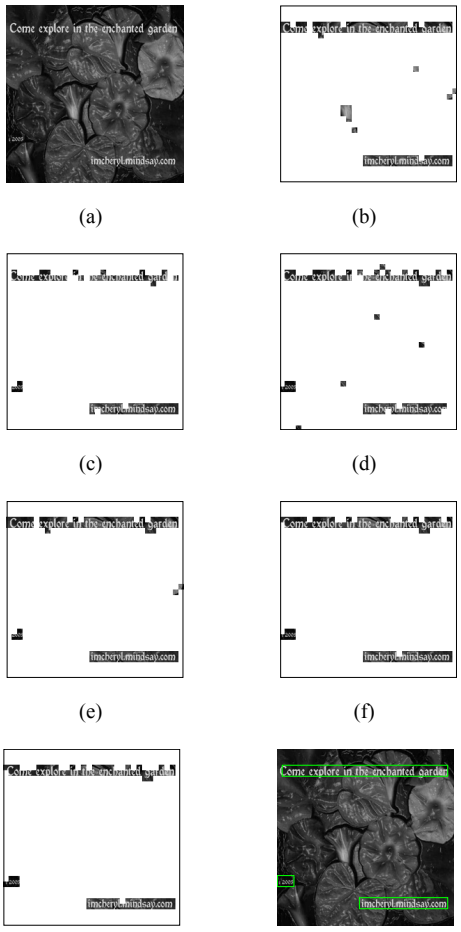
2.2.4. Quadratic Weight

Similarly, for 8×8 DCT coefficients, the quadratic weight function W_Q is given by

$$W_Q(p,q) = \begin{cases} 0, & p = q = 1 \\ (\frac{p+q}{2})^2, & p + q > 2 \end{cases} \tag{10}$$

where quadratic weights further emphasize high frequency components.

Figure 1 shows a detection example using W_{BL}, W_{BM}, W_{BH}, W_U, W_L, and W_Q. In Fig. 1 (b), the W_{BL} detection result loses one text region in left bottom corner and products some false text blocks. Figures 1 (c) and (d) ,W_{BM} and W_{BH}, show better results than Fig. 1 (b), yielding correct text regions and fewer false text blocks. Figure 1 (e), W_U, distinguishes most text blocks with very few false text blocks while Fig. 1 (f) and (g), W_L and W_Q, show completely correct division between text blocks and non-text blocks. In Fig.

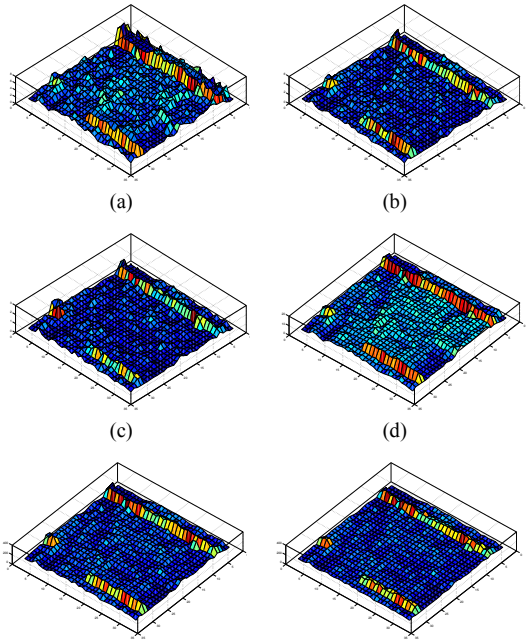


(a) (b) (c) (d) (e) (f) (g) (h)

Fig. 1. (a) input image, (b) detection result with W_{BL}, (c) detection result with W_{BM}, (d) detection result with W_{BH}, (e) detection result with W_U, (f) detection result with W_L, (g) detection result with W_Q, (h) bounding box generated onto the input image.

1 (h), bounding boxes, which are rectangular candidate text regions generated before using an Optical Character Recognition (OCR) engine, are generated based on output shown in Fig. 1 (g). The detailed procedure is as follows. Firstly, the maximum number of blocks horizontally and vertically are calculated. Secondly, these two numbers are the rectangular text region's height and width. Finally, the bounding box is generating along the rectangle contour.

Figure 2 shows energy plots on input image Fig. 1 (a) with different weights applied to the block-wise DCT coefficients. The plots, shown in Fig. 2 (e) and (f), have flatter background and sharper contrast between text regions and background than those shown in Fig. 2 (a)-(d). These observations explain the better performance of the linear and quadratic weights.



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

Fig. 2. (a) energy plot with W_{BL}, (b) energy plot with W_{BM}, (c) energy plot with W_{BH}, (d) energy plot with W_U, (d) energy plot with W_L, (f) energy plot with W_Q.

3. EXPERIMENTAL RESULTS

In this section, the proposed detection algorithms are evaluated through program simulations running on a Microsoft Windows XP, Pentium D, 3.2 GHz platform. The programs are implemented by using Matlab. In the simulation, 20 images with text are randomly selected from our image library. All are gray scale images and of size 256×256. Since the block size does not significantly affect the result, we choose an 8×8 block size. In the following, we compare the performance of proposed algorithms with different types of weights.

We use two metrics (recall and precision) commonly used in Information Retrieval (IR) to evaluate different detection algorithms:

$$\text{recall} = \frac{\text{number of correctly detected text regions}}{\text{number of text regions}}$$

$$\text{precision} = \frac{\text{number of correctly detected text regions}}{\text{number of detected regions}}$$

Table 1 gives the evaluation for different weighted text energy algorithms. The first row is Crandal [3]'s method, which selects 18 coefficients in row-major order as discrimination statistic. Shiratori [4]'s method is equivalent to using binary weight W_{BM}, which is defined in Equation (6). These two methods are evaluated by recall and precision metrics along

Table 1. The evaluation for weighted methods

Weight function	Recall	Precision
Crandall [3]	71.1%	55.2%
W_{BL}	60.7%	52.9%
W_{BM}	78.5%	59.2%
W_{BH}	68.1%	56.1%
W_U	93.3%	58.6%
W_L	95.5%	59.4%
W_Q	96.3%	63.7%

with other weighted energy methods. The table shows that uniform, linear and quadratic weights achieve high recall rate. However, the precision is low for all of them. Further refine operations such as learning machine training can be employed to overcome these problems at additional computational costs. Table 1 shows that the quadratic weight achieve best recall rate 96.3% and precision rate 63.7%.

The evaluations above are all based on the fixed thresholds, which are optimal solutions obtained by maximizing the ratio of between-class scatter to within-class scatter. In some cases, we only pay more attention to accuracy or recall rate, rather than error or false alarm rate, i.e., there is tradeoff between them. To explain this more detail, we introduce the concept of Receiver Operating Characteristics (ROC). In short, a ROC (curve) is a graphical plot of the True Positive Rate (TPR) versus False Positive Rate (FPR) for a binary classifier system as its discrimination threshold is varied. TPR is defined the same as recall while FPR is defined as

$$\text{FPR} = \frac{\text{number of incorrectly detected text regions}}{\text{number of non-text regions}}$$

Figure 3 shows the ROC curves for the different weighted methods. When the FPR is low, the uniform weighted algorithm achieve best TPR. As the demand of TPR increases, linear and quadratic weighted methods achieve the lowest FPR. In other words, linear and quadratic weights bring the benefit of higher TPR with lower FPR. In TPR sensitive cases, linear and quadratic weighted methods are better choices. In FPR sensitive cases, the binary weighted method should be chosen.

4. CONCLUSIONS

In this paper, a weighted DCT-based text energy detection algorithm is proposed for text information extraction. Experimental results show that the quadratic weighted energy method achieves highest recall and precision rates, which is based on fixed optimal threshold generated by LDA. With varied thresholds, linear, quadratic, and binary weighted methods have their advantages in different sensitive cases. Specifically, the quadratic weight should be chosen for the high TPR demand.

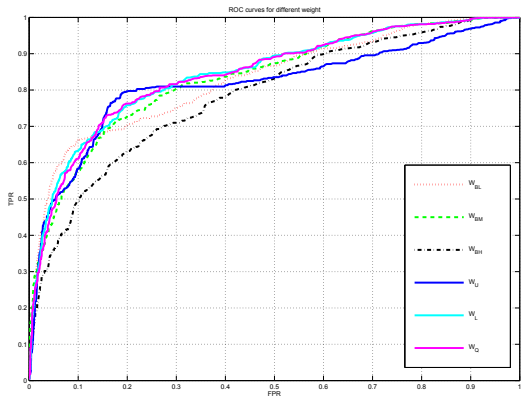


Fig. 3. ROC curves of different weighted methods

5. REFERENCES

[1] N. Ezaki, M. Bulacu, L. Schomaker, "Detection from Natural Scene Image: Towards a System for Visually Impaired Persons," *in Proc. of IEEE International Conference on Pattern Recognition (ICPR'04)*, pp. 683-686, 2004.

[2] W. Wu, X. Chen, J. Yang, "Detection of Text on Road Signs from Video," *IEEE Trans. on Intelligent Transportation Systems*, vol. 6, no. 4, pp. 378-390, Dec. 2005.

[3] D. Crandall, S. Antani, R. Kasturi, "Extraction of special effects caption text events from digital video," *International Journal on Document Analysis and Recognition*, pp. 138-157, 2003.

[4] H. Shiratori, H. Goto, H. Kobayshi, "An Efficient Text Capture Method for Moving Robots Using DCT Feature and Text Tracking," *in Proc. of IEEE International Conference on Pattern Recognition (ICPR'06)*, pp. 1050-1053, 2006.

[5] S. Nadarajah and S. Kotz, "On the DCT Coefficient Distributions," *IEEE Signal Processing Letters*, vol. 13, no. 10, pp. 601-603, Oct. 2006.

[6] E. Y. Lam and J. W. Goodman, "A Mathematical Analysis of the DCT Coefficients," *IEEE Trans. on Image Processing*, vol. 9, no. 10, pp. 1661-1666, Oct. 2000.