European Journal of Scientific Research
ISSN 1450-216X Vol.36 No.4 (2009), pp.502-512
© EuroJournals Publishing, Inc. 2009
http://www.eurojournals.com/ejsr.htm

Document Text Extraction from Document Images Using Haar Discrete Wavelet Transform

S.Audithan
Research Scholar, Dept of CSE
Annamalai University, Annamalai NagarTamil Nadu, India
E-mail: sarabar36@rediffmail.com

RM. Chandrasekaran
Professor, Dept of CSE, Annamalai University, Annamalai Nagar
Tamil Nadu, India
E-mail: aurmc@sify.com

Abstract

This paper presents an efficient and computationally fast method to extract text regions from documents. In this paper, we propose Haar discrete wavelet transform (DWT)[9] which operates the fastest among all wavelets because its coefficients are either 1 or -1. This is one of the reasons we employ Haar DWT to detect edges of candidate text regions. First, we detect edges and then line feature vector graph is generated based on the edge map and the stroke information is extracted. Finally text regions are generated and filtered according to line features. Experimental results show that, without increasing the computational cost, our proposed method could suppress the false alarms notably. Furthermore, our method can be easily customized for applications with different tradeoffs in recall and precision.

1. Introduction

As the requirement of automatic processing on large sum of scanned newspaper images in microfilm format, we demand an automated image processing system to handle the case. Actually the processing of document image segmentation and classification is an OCR pre-processor, as well as pre-processing for understanding document page layout with structured formats. Through this stage page segmentation will need to accurately partition images into blocks of text, figure, table, frame, and other entities[4]. Despite the many efforts spent on the subject there is still much room for improvement in document segmentation techniques, which is the key factor to improve the overall performance of an automatic reading/processing system. Even very good OCR system can be almost useless when text-extraction is performed poorly, which is often the case in existing systems for documents with multiple layouts.

Techniques for document segmentation and layout analysis are traditionally subdivided into three main categories[6]: bottom-up, top-down and hybrid techniques. Some other up-to-date methods are introduced by recent progresses in this area, so as to expand the scope of above categorization. Bottom-up techniques progressively merge evidence at increasing scales to form, e.g., words from characters, lines from words, columns from text lines. They are usually more flexible than top-down methods, but they may suffer from the accumulation of mistakes when going from the small-scale details up to the large scale features. Top-down techniques start by detecting the large-scale features of the image (e.g., columns) and proceed by successive splitting until they reach the smallest-scale features (i.e., individual characters, or text lines). For the procedure to be effective, a priori knowledge about the structure of the page is necessary.

These techniques are therefore particularly useful when the layout is constrained, such as is often the case when considering pages from scientific journals. Most methods do not fit into one of these two categories and are therefore called hybrid. Among these we can find methods based on texture analysis and methods based on background analysis. In methods based on texture analysis the problem of reconstructing the document layout is seen as a problem of texture segmentation. The document page is subdivided into small regions each of which is classified as belonging to one of a few categories (text, drawing, image, etc)[1] according to an analysis of its texture. Once each region in the image has been tentatively classified, a globally consistent segmentation is carried out by the usual techniques of machine vision. Examples of methods using texture analysis are those based on Gabor filtering and mask convolution, fractal signature, and wavelet analysis[9]. All these methods are quite general and flexible but they are also computationally demanding.

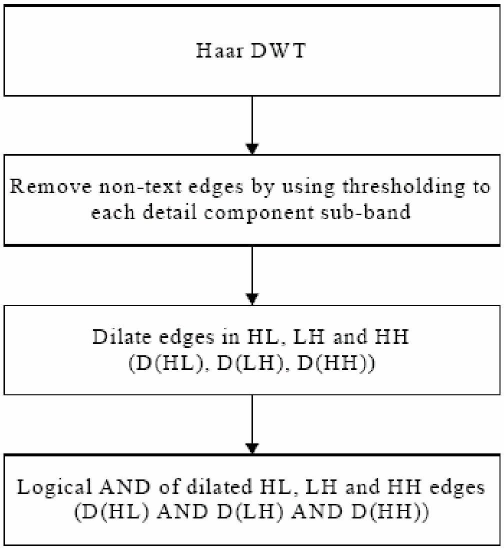
2. DWT Based Extraction

In this section, we present a method to extract texts in document images using Haar discrete wavelet transform (Haar DWT). The edges detection is accomplished by using 2-D Haar DWT[9] and some of the non-text edges are removed using thresholding. Afterward, we use different morphological dilation operators to connect the isolated candidate text edges in each detail component sub-band of the binary image. Although the color component may differ in a text region, the information about colors does not help extracting texts from images. If the input image is a gray-level image, the image is processed directly starting at discrete wavelet transform[2]. If the input image is colored, its RGB components are combined to give an intensity image Y as follows:

$$Y=0.299\ 0.587\ 0.114\ \mathit{Y\ R\ G} \qquad (1)$$

Image Y is then processed with discrete wavelet transform and the whole extraction algorithm afterward. If the input image itself is stored in the DWT compressed form, DWT operation can be omitted in the proposed algorithm. The flow chart of the proposed algorithm is shown in Figure 1.

Figure 1: Flow chart of the proposed text extraction algorithm


Haar DWT
Remove non-text edges by using thresholding to each detail component sub-band
Dilate edges in HL, LH and HH (D(HL), D(LH), D(HH))
Logical AND of dilated HL, LH and HH edges (D(HL) AND D(LH) AND D(HH))

2.1. Haar discrete wavelet transform

The discrete wavelet transform is a very useful tool for signal analysis and image processing, especially in multi-resolution representation. It can decompose signal into different components in the frequency domain. One-dimensional discrete wavelet transform (1-D DWT) decomposes an input sequence into two components (the average component and the detail component) by calculations with a low-pass filter and a high-pass filter. Two-dimensional discrete wavelet transform (2-D DWT) decomposes an input image into four sub-bands, one average component (LL)[7] and three detail components (LH, HL, HH) as shown in Figure 2. In image processing, the multi-resolution of 2-D DWT has been employed to detect edges of an original image. The traditional edge detection filters can provide the similar result as well. However, 2-D DWT can detect three kinds of edges at a time while traditional edge detection filters cannot. As shown in Figure 3, the traditional edge detection filters detect three kinds of edges by using four kinds of mask operators. Therefore, processing times of the traditional edge detection filters is slower than 2-D DWT.

Figure 2: The result of 2-D DWT decomposition

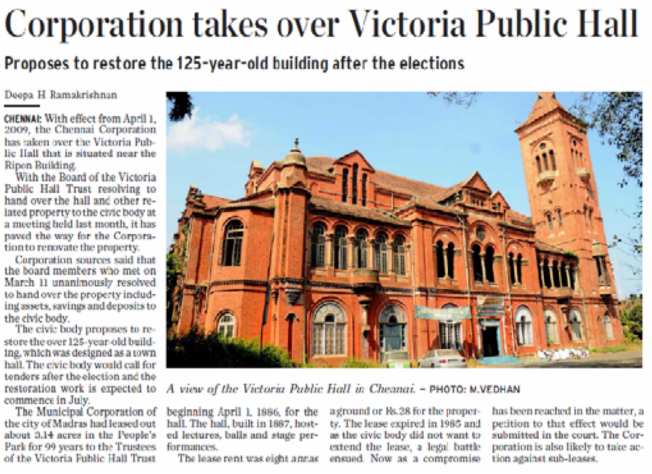
LL	HL
LH	HH

Three kinds of edges present in the detail component sub-bands but look unobvious (very small coefficients). DWT filters with Haar DWT, the detected edges become more obvious and the processing time decreases. The operation for Haar DWT is simpler than that of any other wavelets. It has been applied to image processing especially in multi-resolution representation.

Harr DWT has the following important features

1. Haar wavelets are real, orthogonal, and symmetric.
2. Its boundary conditions are the simplest among all wavelet-based methods.
3. The minimum support property allows arbitrary spatial grid intervals.
4. It can be used to analyze texture and detect edges of characters.
5. The high-pass filter and the low-pass filter coefficient is simple (either 1 or –1).

Figure 3 (a): Original RGB Image



Corporation takes over Victoria Public Hall

Proposes to restore the 125-year-old building after the elections

Deepa H Ramakrishnan

CHENNAI: With effect from April 1, 2009, the Chennai Corporation has taken over the Victoria Public Hall that is situated near the Ripon Building.

With the Board of the Victoria Public Hall Trust resolving to hand over the hall and other related property to the civic body at a meeting held last month, it has paved the way for the Corporation to renovate the property.

Corporation sources said that the board members who met on March 11 unanimously resolved to hand over the property including assets, savings and deposits to the civic body.

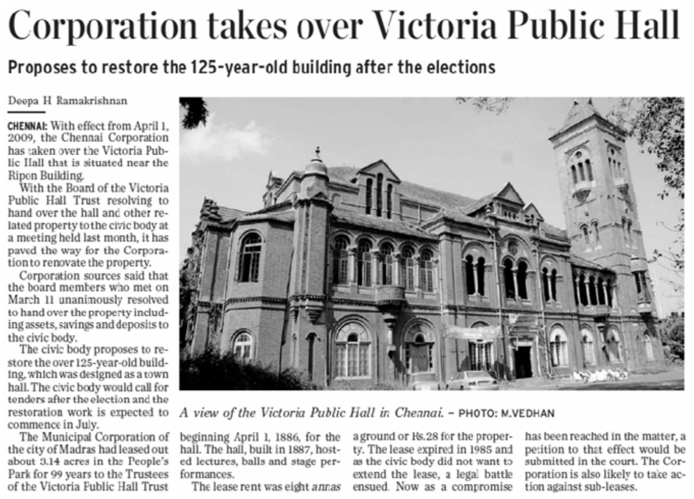
The civic body proposes to restore the over 125-year-old building, which was designed as a town hall. The civic body would call for tenders after the election and the restoration work is expected to commence in July.

The Municipal Corporation of the city of Madras had leased out about 3.14 acres in the People's Park for 99 years to the Trustees of the Victoria Public Hall Trust beginning April 1, 1886, for the hall. The hall, built in 1887, hosted lectures, balls and stage performances.

The lease rent was eight annas aground or Rs.28 for the property. The lease expired in 1985 and as the civic body did not want to extend the lease, a legal battle ensued. Now as a compromise has been reached in the matter, a petition to that effect would be submitted in the court. The Corporation is also likely to take action against sub-leases.

A view of the Victoria Public Hall in Chennai. – PHOTO: M.VEDHAN

(b): Grayscale Image



Corporation takes over Victoria Public Hall

Proposes to restore the 125-year-old building after the elections

Deepa H Ramakrishnan

CHENNAI: With effect from April 1, 2009, the Chennai Corporation has taken over the Victoria Public Hall that is situated near the Ripon Building.

With the Board of the Victoria Public Hall Trust resolving to hand over the hall and other related property to the civic body at a meeting held last month, it has paved the way for the Corporation to renovate the property.

Corporation sources said that the board members who met on March 11 unanimously resolved to hand over the property including assets, savings and deposits to the civic body.

The civic body proposes to restore the over 125-year-old building, which was designed as a town hall. The civic body would call for tenders after the election and the restoration work is expected to commence in July.

The Municipal Corporation of the city of Madras had leased out about 3.14 acres in the People's Park for 99 years to the Trustees of the Victoria Public Hall Trust beginning April 1, 1886, for the hall. The hall, built in 1887, hosted lectures, balls and stage performances.

The lease rent was eight *annas* a ground or Rs.28 for the property. The lease expired in 1985 and as the civic body did not want to extend the lease, a legal battle ensued. Now as a compromise has been reached in the matter, a petition to that effect would be submitted in the court. The Corporation is also likely to take action against sub-leases.

A view of the Victoria Public Hall in Chennai. – PHOTO: M.VEDHAN

(c): Binary Image

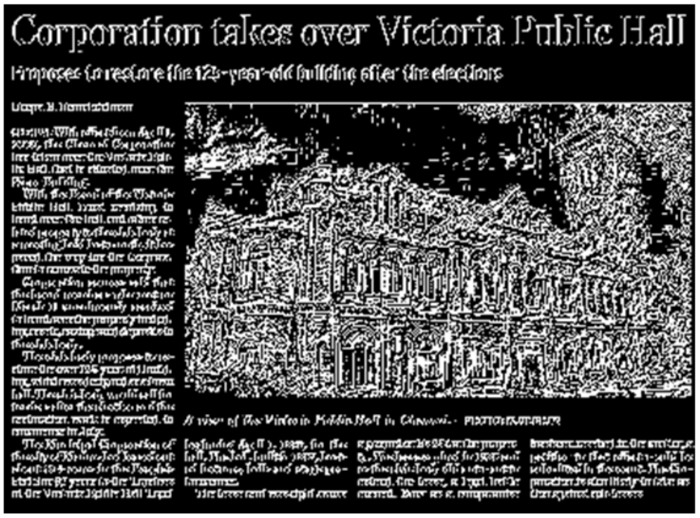


Figure 3 (a) shows the example of a 4×4 gray-level image. The wavelet coefficients can be obtained in gray-level image using addition and subtraction. 2-D DWT is achieved by two ordered 1-D DWT operations (row and column). First of all, we perform the row operation to obtain the result shown in Figure 3(b). Then it is transformed by the column operation and the final resulted 2-D Haar DWT is shown in Figure 3 (c). 2-D Haar DWT decomposes a gray-level image into one average component sub-band and three detail component sub-bands.

Figure 4 (a): The original image (b) the row operation of 2-D Haar DWT (c) the column operation of 2-D Haar DWT

$$\begin{bmatrix} A & B & C & D \\ E & F & G & H \\ I & J & K & L \\ M & N & O & P \end{bmatrix} \qquad \begin{bmatrix} (A+B) & (C+D) & (A-B) & (C-D) \\ (E+F) & (G+H) & (E-F) & (G-H) \\ (I+J) & (K+L) & (I-J) & (K-L) \\ (M+N) & (O+P) & (M-N) & (O-P) \end{bmatrix}$$

(a) (b)

$$\begin{bmatrix} (A+B)+(E+F) & (C+D)+(G+H) & (A-B)+(E-F) & (C-D)+(G-H) \\ (I+J)+(M+N) & (K+L)+(O+P) & (I-J)+(M-N) & (K-L)+(O-P) \\ (A+B)-(E+F) & (C+D)-(G+H) & (A-B)-(E-F) & (C-D)-(G-H) \\ (I+J)-(M+N) & (K+L)-(O+P) & (I-J)-(M-N) & (K-L)-(O-P) \end{bmatrix}$$

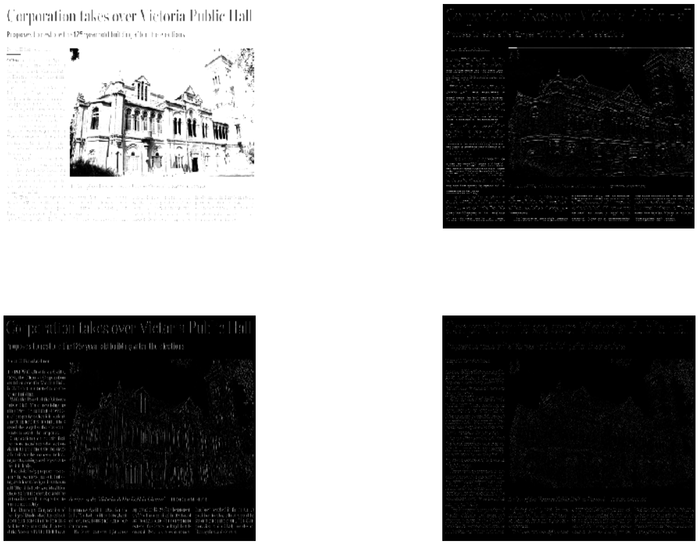
(c)

In those three detail components of a Haar DWT image, we can obtain various features about the original image as follows:

1. Average components are detected by the LL sub-band;
2. Vertical edges are detected by the HLsub-band;
3. Horizontal edges are detected by the LH sub-band;
4. Diagonal edges are detected by the HH sub-band.

For example, the gray-level image shown in Figure 4 (a) is decomposed into 2-D Haar DWT as shown in Figure 5. We can detect candidate text edges in the original image from those three detail component sub-bands (HL, LH and HH) in Figure 5.

Figure 5: 2-D Haar discrete wavelet transform image



Chen and Liao presented the segment-matrix algorithm for Haar DWT to decrease the processing time of DWT operations. The method produces the same results as traditional Haar DWT with a much faster speed. Hence, we apply the segment-matrix algorithm to decompose an original gray-level image [3] into four sub-bands. Figure 4 (a) shows an example of a 4×4 gray-level image. It is segmented into 4 2×2 sub-blocks as shown in Figure 4 (b). Then each 2×2 sub-block is performed with the z-scan operation and we Obtain 4 1×4 sub-blocks as shown in Figure 4 (c). The Haar DWT filter coefficient matrix (presented in Figure 6(d)) is multiplied by the matrix shown in Figure 4 (c) and then the result of 2-D DWT is obtained in Figure 6 (e). After the Haar DWT, the detected edges include mostly text edges and some non-text edges are presented in the 3 detail component sub-bands. In next subsection, we employ dynamic thresholding to preliminarily remove those non-text edges in the detail component sub-bands.

Figure 6: (a) The original image (b) the original image to change into 2×2 sub-blocks (c) the z-scan result of the 2×2 sub-blocks (d) the filter coefficient matrix (e) the correct result of 2-D Haar DWT

$$\begin{bmatrix} A & B & C & D \\ E & F & G & H \\ I & J & K & L \\ M & N & O & P \end{bmatrix}$$

(a)

$$\begin{bmatrix} \begin{bmatrix} A & B \\ E & F \end{bmatrix} & \begin{bmatrix} C & D \\ G & H \end{bmatrix} \\ \begin{bmatrix} I & J \\ M & N \end{bmatrix} & \begin{bmatrix} K & L \\ O & P \end{bmatrix} \end{bmatrix}$$

(b)

$$\begin{bmatrix} [A \quad B \quad E \quad F] \\ [C \quad D \quad G \quad H] \\ [I \quad J \quad M \quad N] \\ [K \quad L \quad O \quad P] \end{bmatrix}$$

(c)

$$\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix}$$

(d)

$$\begin{bmatrix} A+B+E+F & C+D+G+H & A-B+E-F & C-D+G-H \\ I+J+M+N & K+L+O+P & I-J+M-N & K-L+O-P \\ A+B-E+F & C+D-G+H & A-B-E-F & C-D-G-H \\ I+J-M+N & K+L-O+P & I-J-M-N & K-L-O-P \end{bmatrix}$$

(e)

2.2. Thresholding

Thresholding is a simple technique for image segmentation [8]. It distinguishes the image regions as objects or the background. Although the detected edges are consist of text edges and non-text edges in every detail component sub-band, we can distinguish them due to the fact that the intensity of the text edges is higher than that of the non-text edges. Thus, we can select an appropriate threshold value and preliminarily remove the non-text edges in the detail component sub-bands. In this subsection, we employ dynamic thresholding to calculate the target threshold value *T*. The target threshold value is obtained by performing an equation on each pixel with its neighboring pixels. We employ two mask operators to obtain such an equation and then calculate the threshold value for each pixel in the 3 detail sub-bands. Basically, the dynamic thresholding method obtains different target threshold values for different images. Each detail component sub-band ***es*** is then compared with *T* to obtain a binary image (***e***).

The threshold *T* is determined by

$$T = \frac{\sum (es(i,j) \times s(i,j))}{\sum s(i,j)} \tag{2}$$

$$s(i,j) = Max(\left|g1 ** es(i,j)\right|, \left|g2 ** es(i,j)\right|) \tag{3}$$

And

$$g1=\begin{bmatrix}-1 & 0 & 1\end{bmatrix}, g2=\begin{bmatrix}-1 & 0 & 1\end{bmatrix}^t \tag{4}$$

In Eq. (3), "**" denote two-dimensional liner convolution. Figure 7 shows the example of a 5×5 detail component sub-band ***(es)***. We calculate S (P8) as an example to demonstrate the definition of Eqs. (3) and (4).

$$S(P8) = max\left(\left|P9 - P7\right|, \left|P13 - P3\right|\right) \tag{5}$$

Applying similar operations to each pixel, we obtain all the S(i, j) for each detail component sub-band. After that, we can apply Eq. (2) to compute *T* and the binary edge image (***e***) is then given by

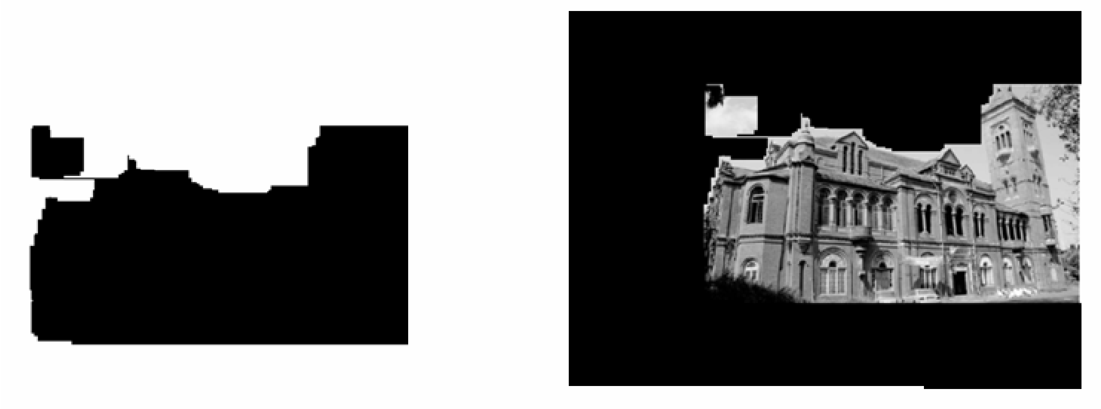
$$e(i,j)=\begin{cases}255, & if\ es(i,j)>T\\ 0\ , & otherwise\end{cases} \tag{6}$$

The resulted binary image, as shown in Figure 8 , is mostly consisted of text edges and very few non-text edges.

Figure 7: 5×5 detail component sub-band *(es)*

$$\begin{bmatrix} P1 & P2 & P3 & P4 & P5 \\ P6 & P7 & P8 & P9 & P10 \\ P11 & P12 & P13 & P14 & P15 \\ P16 & P17 & P18 & P19 & P20 \\ P21 & P22 & P23 & P24 & P25 \end{bmatrix}$$

Figure 8: Binary image of detail component sub-band



2.3. Text region extraction

In this subsection, we use morphological operators and the logical AND operator to further removes the non-text regions. In text regions, vertical edges, horizontal edges and diagonal edges are mingled together while they are distributed separately in non-text regions. Since text regions are composed of vertical edges, horizontal edges and diagonal edges, we can determine the text regions to be the regions where those three kinds of edges are intermixed. Text edges are generally short and connected with each other in different orientation. In Figure 9, we use different morphological dilation [11] operators to connect isolated candidate text edges in each detail component sub-band of the binary image. In this paper, 3×5 for horizontal operators, 3×3 for diagonal operators and 7×3 for vertical operators as in shown Figure 10 are applied. The dilation operators for the three detail sub-bands are designed differently so as to fit the text characteristics. The logical AND is then carried on three kinds (vertical, horizontal and diagonal) of edges after morphological dilation. This process is indicated in Figure 11. Since three kinds of edge regions are intermixed in the text regions, overlapping appears a lot after the morphological dilation due to the expansion of each single edge. On the contrary, only one kind of edge region or two kinds of edge regions exist separately in the non-text regions and hence there is no overlapping even after the dilation. Therefore, the AND operator helps us to obtain the candidate text regions as shown in Figure 12. Sometimes the text candidate regions may contain some non-text component regions which are too large or too small. By limiting the block size, we obtain the final text regions. Each text region has a moderate size w × h (pixels) in a candidate text region image.

Figure 9: Image after morphological Operations

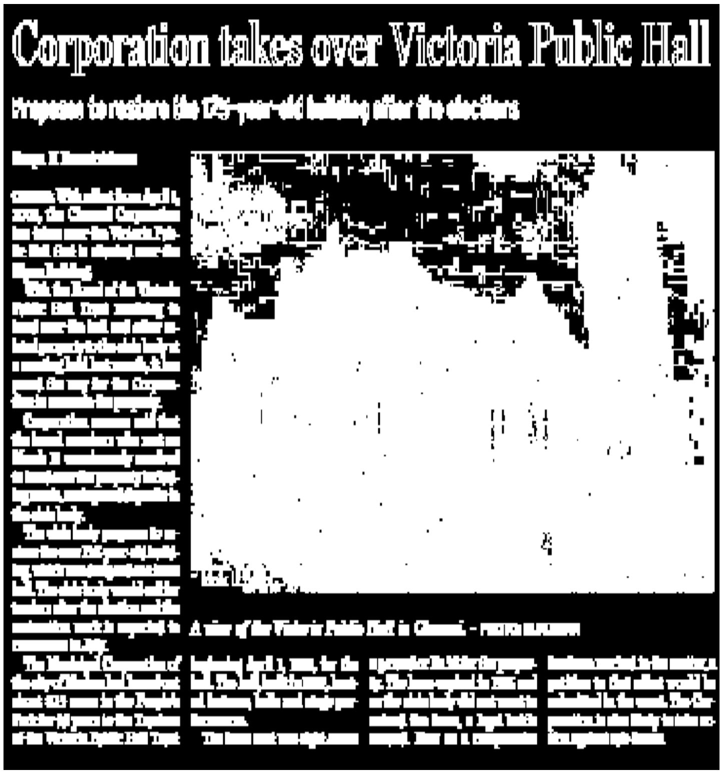


Figure 10: Horizontal, Diagonal and Vertical edges dilation operators

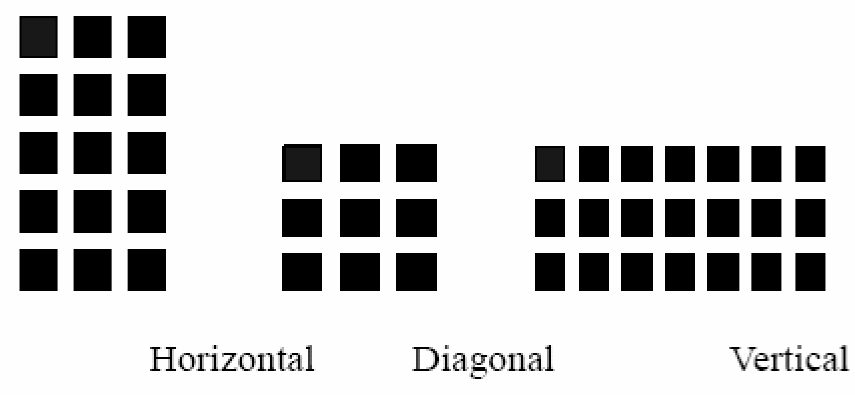
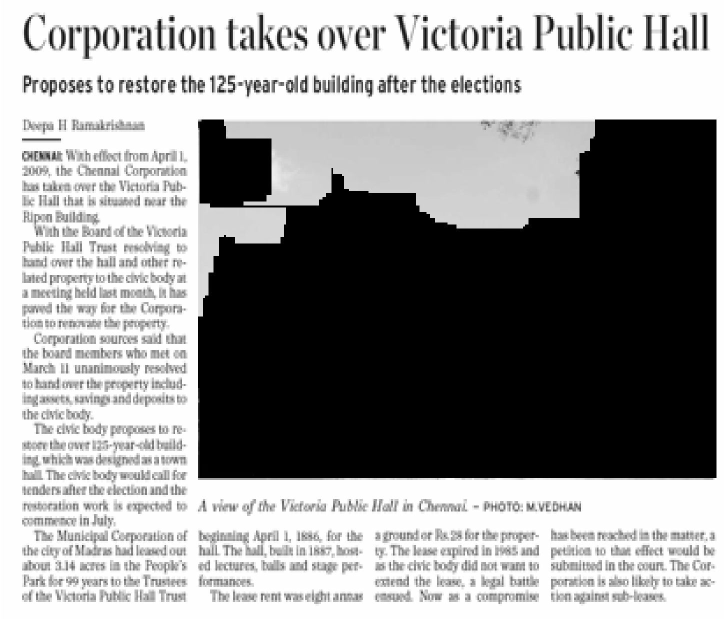


Figure 11: Original image



Corporation takes over Victoria Public Hall

Proposes to restore the 125-year-old building after the elections

Deepa H Ramakrishnan

CHENNAI: With effect from April 1, 2009, the Chennai Corporation has taken over the Victoria Public Hall that is situated near the Ripon Building.

With the Board of the Victoria Public Hall Trust resolving to hand over the hall and other related property to the civic body at a meeting held last month, it has paved the way for the Corporation to renovate the property.

Corporation sources said that the board members who met on March 11 unanimously resolved to hand over the property including assets, savings and deposits to the civic body.

The civic body proposes to restore the over 125-year-old building, which was designed as a town hall. The civic body would call for tenders after the election and the restoration work is expected to commence in July.

A view of the Victoria Public Hall in Chennai. – PHOTO: M.VEDHAN

The Municipal Corporation of the city of Madras had leased out about 3.14 acres in the People's Park for 99 years to the Trustees of the Victoria Public Hall Trust beginning April 1, 1886, for the hall. The hall, built in 1887, hosted lectures, balls and stage performances.

The lease rent was eight annas a ground or Rs.28 for the property. The lease expired in 1985 and as the civic body did not want to extend the lease, a legal battle ensued. Now as a compromise has been reached in the matter, a petition to that effect would be submitted in the court. The Corporation is also likely to take action against sub-leases.

The minimum text block size is determined as follows:

Width>100 (pixels), height >35(pixel)

Removing the candidate text regions smaller than this limit, the final text region is shown in Figure 12.

3. Experimental Result

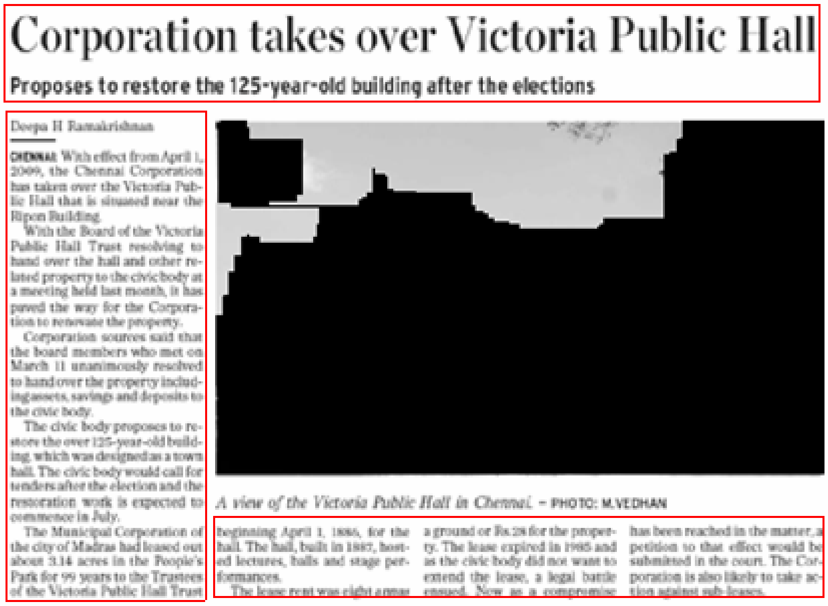
We conducted experiments on the PC platform of Pentium IV at 1.8GHz. The experimental data are various documents of different content and the ground truth of the texts in the frames were obtained manually. There are 250 files containing 1696 text regions. Two different types of performance measure are used to evaluate our text detection method: box -based measure and pixel –based measure.

The experimental results show that our proposed detection method can achieve 94.8% detection rate and 7.1% false alarm rate. To verify the effectiveness of this method, we have implemented a typical edge–based method proposed. This algorithm computes maximum gradient difference to detect potential text line segment from horizontal scan lines of the document. The comparison experiment results were based on pixel measure.

The comparison experiment results show that without increasing the processing time, the performance of our method is superior to the MGD method both in precision and in false alarm rate. This is because our method exploits the improved canny edge detector and takes use of the line feature of characters.

Furthermore, a threshold can be set for different values according to different requirements. Specifically, in some applications where higher detection rate is preferred, one –phase thresholding with a moderate value can be used. In other applications where higher accuracy is preferred, lower false alarm rate is preferred; the two-phase thresholding with a higher value can meet such requirements.

Figure 12: Document Text Extraction



Corporation takes over Victoria Public Hall

Proposes to restore the 125-year-old building after the elections

Deepa H Ramakrishnan

CHENNAI: With effect from April 1, 2009, the Chennai Corporation has taken over the Victoria Public Hall that is situated near the Ripon Building.

With the Board of the Victoria Public Hall Trust resolving to hand over the hall and other related property to the civic body at a meeting held last month, it has paved the way for the Corporation to renovate the property.

Corporation sources said that the board members who met on March 11 unanimously resolved to hand over the property including assets, savings and deposits to the civic body.

The civic body proposes to restore the over 125-year-old building, which was designed as a town hall. The civic body would call for tenders after the election and the restoration work is expected to commence in July.

The Municipal Corporation of the city of Madras had leased out about 3.14 acres in the People's Park for 99 years to the Trustees of the Victoria Public Hall Trust beginning April 1, 1886, for the hall. The hall, built in 1887, hosted lectures, balls and stage performances.

The lease rent was eight annas a ground or Rs.28 for the property. The lease expired in 1985 and as the civic body did not want to extend the lease, a legal battle ensued. Now as a compromise has been reached in the matter, a petition to that effect would be submitted in the court. The Corporation is also likely to take action against sub-leases.

A view of the Victoria Public Hall in Chennai. – PHOTO: M.VEDHAN

Fig 12 gives some more results of text detection. The highlight rectangles are the bounding boxes of the detected texts. In fig both high-contrast (in the lower-right of the image) and low-contrast (in the middle of the image) are successfully detected. In fig besides text lines superimposed on simple background (the black stripe background) are detected, text strings embedded in complex background (the grass background) are also detected. The vertical lines are missed because only horizontal text lines are considered.

4. Conclusion

In this paper, we propose an effective document text extraction method based on line features. This method exploits an improved canny edge detector to detect text pixels. By considering the spatial distribution of edge pixels, stroke information is incorporated. Results show that our method can obtain 95% hit rate and only 7% false alarm rate, which is superior to the typical edge-based text extraction methods without increasing the computational cost.

References

[1] Park, C. J., Moon, K. A., Oh, Weon-Geun, and Choi, H. M. 2000. An efficient of character string positions using morphological operator. IEEE International Conferences on Systems, Man, and Cybernetics, 3, 8-11: 1616-1620.

[2] Zhong, Yu., Karu, K., and Jain, A.K.1995. Locating text in complex color images. Proceedings of the Third International Conference on Document Analysis and Recognition, 1, 14-16:146-149.

[3] Chen, Datong, Bourlard, H., and Thiran J. P., 2001. Text identification in complex background using SVM. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Proceedings of the 2001, 2, 8-14: 621-626.

[4] Lam, S. W., Wang, D., and Srihari, S. N., 1990. Reading newspaper text. International Conference on Pattern Recognition Proceedings, 10th. I, 16-21:

[5] Williams, P. S. and Alder, M. D. 1996.Generic texture analysis applied to newspaper egmentation. IEEE International Conference on Neural Networks, 3, 3-6: 1664-1669.

[6] Zhong, Yu. Zhang, Hongjiang. And Jain, A. K. 2000. Automatic caption localization in compressed video. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22, 4: 385 –392.

[7] Acharyya, M. and Kundu, M. K. 2002.Document image segmentation using wavelet scale-space features. IEEE Transactions on Circuits and Systems for Video Technology, 12, 12: 1117 –1127.

[8] Mallat, S. G. 1989. A theory for multiresolution signal decomposition: the wavelet representation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 11, 7: 674-693.

[9] Acharya, Tinku., Chen, Po-Yueh. 1998. VLSI implementation of DWT architecture. Proceedings of the IEEE International Symposium on Circuits and Systems, 2: 272-275.

[10] Grochening, K. and Madych, W. R. 1992. Multiresoultion analysis, Haar bases, and self-similar tilings of Rn. IEEE Transactions on Information Theory, 38, 2.

[11] Fujii, Masafumi., Wolfgang, J. R., and Hoefer. 2001. Filed-Singularity correction in 2-D time-domain Haar-wavelet modeling of waveguide components. IEEE Transactions on Microwave Theory and Techniques, 49, 4.

[12] Chen, P. Y. and Liao, E. C. 2002. A new algorithm for Haar discrete wavelet transform.IEEE International Symposium on Intelligent Signal Processing and Communication Systems, 21, 24: 453-457.

[13] Hasan, M. Y. and Lina J, Y. K. Morphological text extraction from image. IEEE Transactions on Image Processing, 9, 11: 1978 -1983.