Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 2, Issue. 6, June 2013, pg.285 – 293

RESEARCH ARTICLE

Multioriented and Curved Text Lines Extraction from Documents

Vaibhav Gavali[1], B. R. Bombade[2]
[1]M. Tech Student, Department of CSE, SGGS IE&T, Nanded, India
[2]Assistant professor, Department of CSE, SGGS IE&T, Nanded, India

[1] *vaibhavkgavali@gmail.com;* [2] *b.r.bombade@gmail.com*

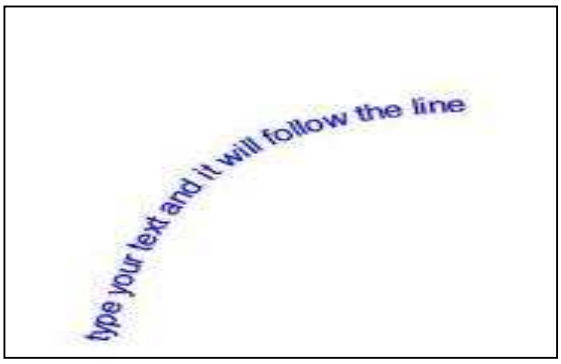
Abstract— There is need of the robust algorithm to extract text lines from script independent documents, color independent, font and size independent segmentation algorithms. This paper presents simple method to extract curved and multioriented text lines from the documents. The input is may be colored or grayscale image. Discrete wavelet transform is applied on input image to get four sub-bands. Thresholding is applied on the three sub-bands (horizontal, vertical, diagonal). Edge detection is applied on the each sub-band after thresholding. Morphological dilation using different dilation operators for different sub-bands are used. Finally logical and operator is on the 3 dilated components to get the curved and Multioriented text lines as output.

Key Terms: - Curved; edge; text lines; dilation; wavelet transform

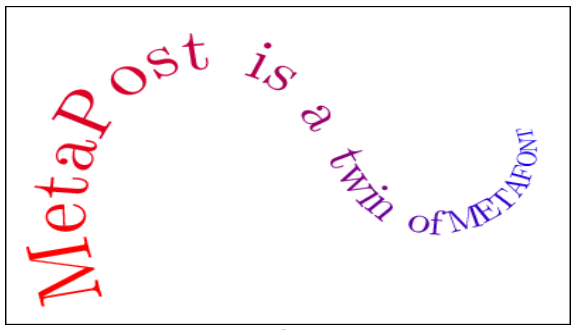
I. INTRODUCTION

Text detection or extraction from the documents is the first step before segmentation. There are many artistic documents such as advertisement, poster, or graphic illustrative where text lines are not single oriented. These text lines may be Multioriented or curved in shape. Examples of such documents are shown in fig. 1. There are many algorithms proposed for straight lines, parallel lines to each other like X-Y cut algorithm, projection profile algorithm, RLSA algorithm etc. But curvilinear, skewed and Multioriented text lines containing documents is a challenge to algorithms developed for machine printed documents [9].

For document segmentation there are two main approaches are used for text extraction from document images those are region based and texture based approach. Region based approach uses properties of color text region and the difference between the background and text region since text regions have higher horizontal spatial variance than that of non-text regions [2]. It is subdivided into two approaches edge based and connected components based methods. Texture based approach it uses distinct textural properties to distinguish between background and text regions [10]. It is highly complex in nature but more robust the connected components methods. Examples of methods using texture analysis are those based on Gabor filtering and mask convolution, fractal signature and wavelet analysis etc. Haar Wavelet transform is computationally demanding since coefficients of wavelet transform are either 1 or -1.



(a)



(b)

(c)

Fig.1. (a) Curved text, (b) Curved and Multioriented text, (c) Skewed Multioriented text.

These wavelets are real, orthogonal symmetric and it allows perfect localization of transform domain [1]. An application of the wavelet transform is the formation of the classification feature using the statistical characteristics of wavelet coefficients [5].

Detection and extraction of text from document images involves challenges since the characters may be of any color, background color, different color intensity, font size and font style of the characters [5]. Luminance of the document may vary due to degradation and presence of noise

There are wide areas and fields like postal code from address on the envelopes and sorting of mail, bank cheque processing, in libraries for computerized the storage of book and texts, also reading devices for blind people etc.[2] where document segmentation is required.

II. RELATED WORK

In Adaptive Color Reduction (ACR) technique and Page Layout Analysis (PLA) approach ACR technique is used to obtain optimal number of colors and convert the document into principal colors using color quantization

algorithm [6]. Then the document image is split into the separable color plains. PLA is applied to each color plains and identifies the text regions. Merging procedure is applied to merge the text regions derived from the color plains and to produce output document [6]. Using contourlet transform image is transformed into directional sub-bands with texture details. Contourlet transform does not offer a high degree of directionality and anisotropy than wavelet transform [7]. Since it is not shift variant NSCT (non sub-sampled contourlet transform) is used. It captures Multioriented texture details at high frequency components so as to produce text regions whereas at low frequency gives rise to non-text region [7].

Using wavelet transform a method is proposed for text extraction from different kinds of images (document image, scene text image, caption text image) [1]. A proposed method for kannada text extraction from images/videos from documents. It uses color reduction technique and standard deviation based technique to detect edges and localization of text regions [4]. Another method for text extraction from the color images using wavelet transform for documents may contain different objects with text, any color, background of the document may be different [5]. A system is proposed for text extraction and removal of non-text areas using wavelet transform in color images from static or video sequences [2].Text extraction algorithm is proposed to extract text from newspaper using wavelet transform [3].

A method proposed for curved and multioriented text lines extraction based on the concept of water reservoir analogy. A reservoir is a metaphor to illustrate the cavity region of a character where water can be stored [8]. In this method first connected components are labeled and identified either as isolated touching. Then each touching component is classified either straight type or curve type depending on reservoir base area and envelope points. Based on the type of the component two candidate points are computed from each touching component. Finally candidate points of each component are detected and after analogy these candidate regions components are grouped to get individual text lines [8]. Local linearity based technique to identify Multioriented or curved text lines from English and Chinese document [11]. First it split the document image into some small constant width and then local orientation is estimated in each of these sub-regions. Extract text lines by extending local orientation of the sub-regions.

Different methods for extraction of text line in arbitrary orientations from mixed text and graphics regions. It uses character bounding box and hough transform but cannot handle curved text lines and lines of arbitrary size [12]. In fuzzy curve-tracing algorithm to detect curve text lines from English and Chinese documents. Here, at first, character pixels are grouped based n the fuzzy c-means algorithm. Each cluster center represents all its associated pixels in a class to reduce the amount of data substantially. Then, analyzing the spatial relationship, cluster centers are connected to generate the initial curve representing the text path. Finally, the text pixels are clustered again under the constraint that the path passing through the cluster centers must be smooth. Main drawback of the method is to define the smoothness term properly [13]. Method proposed to extract straight text lines in arbitrary orientations. In this method, line anchors are first found in the document image and then text lines are generated expanding the line anchors [14].

A computational geometric approach to extract text lines from a document. Although this method can handle document images with complex layout, it cannot handle curved text lines [15].

III. PROPOSED ALGORITHM

We present a method for extraction of text lines from the curved and multioriented text documents using the Haar discrete wavelet transform (Haar DWT). The edges detection is performed by using 2-D Haar DWT [3] 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 then its RGB components are used to get intensity of the image as follows

$$Y=0.299R+0.587G+0.114B$$

If the input image itself is stored in the DWT compressed form, DWT operation can be omitted in the proposed algorithm. The following flow chart describes the complete algorithm [5] as shown in fig. 2.

A. Haar Discrete Wavelet Transform

It is used to get frequency components of the document. It uses two filters for two times to get more combinations of the filters to filter for getting components which are low pass and high pass filter.

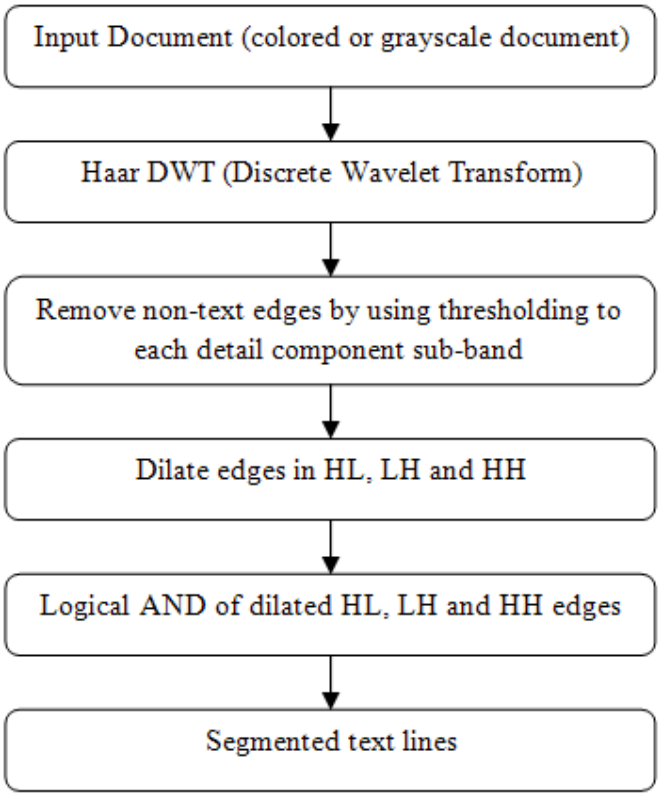


Fig.2. Proposed text line extraction algorithm

It decomposes image into four components using two filters LL, LH, HL, HH. Out of which LL is an average component derived and LH, HL, HH are detail components as shown in the fig. 3. 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 [5]. Since 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 [5].

LL	**HL**
LH	**HH**

Fig.3. Sub-bands after DWT decomposition

B. Thresholding

Thresholding is an important for document segmentation. It distinguishes the image regions as objects or the background. Although the detected edges are consist of text edges and non-text edges or background 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 as background. Thus, we can select an appropriate threshold value and preliminarily separate background from text region 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***) [3].

The threshold T is determined by

$$T = \frac{\Sigma(es(i,j)\times s(i,j))}{\Sigma(s(i,j))} \quad (1)$$

Where

$$S(i,j)=Max(|g1**es(i,j)|,|g2**es(i,j)|) \quad (2)$$

and

$$g1=[-1\ 0\ 1],\ g2=[-1\ 0\ 1]' \quad (3)$$

In Eq. (2), "**" denote two-dimensional liner convolution.

After calculating s(i, j) for each detail component sub-band we can apply Eq. (1) to compute *T* and the binary edge image (***e***) is then given by

$$e(i,j)=\begin{cases}255, if\ es(i,j) > T\\ 0, otherwise\end{cases} \quad (4)$$

In e image we get most of the text edges. It is obtained for each of the sub-band separately [3].

C. Text Region Extraction

In this subsection we apply morphological dilation operation and logical AND operator for text extraction. We use different morphological operators on each of the sub-band to get the text regions. In text regions, vertical edges, horizontal edges and diagonal edges are mingled together while they are distributed separately in 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 [5]. Text edges are generally short and connected with each other in different orientation. We use 3×5 for horizontal operators, 3×3 for diagonal operators and 7×3 for vertical operators as shown in fig.4. 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 [5]. 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 background and hence there is no overlapping even after the dilation. Therefore, the AND operator helps us to obtain the candidate text regions. Sometimes the text candidate regions may contain some non-text or background component regions which are too large or too small [5].

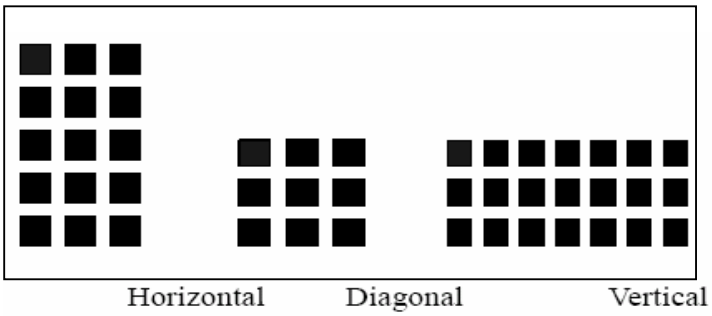


Fig.4. Different morphological dilation operators

After the logical AND operation performed on the dilated edges of the sub-bands we get the text lines extracted from the documents.

IV. EXPERIMENTAL RESULTS

For experimental purpose we have chosen 100 sample document images of different size of pixels. Out of which 20 curved documents and 80 multioriented and skewed documents. Dataset is chosen from different posters, covers of book, greeting cards, advertisements, magazine covers, newspaper cuttings, graffiti phrases from devnagari and English scripts.

The overall processing cost of the method depends on the size of the input document image and the size of the obtained text regions according to document size.

The performance was evaluated by comparing the manually extracted ground truth regions with the automatically extracted regions. The precision and recall rates (Equations (5) and (6)), have been computed based on the number of correctly detected words in document image in order to further evaluated the efficiency and robustness. The precision rate is defined as the ration of correctly detected words to the sum of correctly detected words plus false positive. False positive are those regions in the image, which are actually not characters of text, but have detected by the algorithm as text regions.

$$\text{Precision Rate} = \frac{\text{Correctly detected words}}{\text{Correctly detected words + False Positives}} * 100\% \quad (5)$$

The Recall rate is defined as the ratio of correctly detected words to the sum of correctly detected words plus false negatives. False negatives are those regions in the image which are actually text characters, but have been not detected by the method.

$$\text{Recall Rate} = \frac{\text{Correctly detected words}}{\text{Correctly detected words + False Negatives}} * 100\% \quad (6)$$

Figures 5(a), 5(c), 5(e), 5(g), 5(i) are the input given to our method and figures 5(b), 5(d), 5(f), 5(h), 5(j) are the output documents of given documents respectively.

TEST DATA	NO OF DOCUMENTS	PRECISION RATE	RECALL RATE
CURVED DOCUMENTS	20	92.4	91.4
MULTIORIENTED DOCUMENTS	80	88.6	93.2
TOTAL	100	90.5	92.3

Table 1. Analysis of precession rate and recall rate

We have obtained 90.5% precision rate for our method and 92.25% recall rate for our method using total dataset of curved and multioriented document images which is shown in the table 1. The percentage of precision and recall rate calculated is optimal due to small and single characters present in our documents. In case of the distance between two words is high then it is difficult to detect those lines for water reservoir technique which is removed by using our method.

A threshold can be used for different values according to the different requirements in aspects of accuracy and text detection rate from the document. In case of some applications where higher detection rate is preferred, one-phase thresholding with moderate value can be used. In other applications where higher accuracy is preferred, lower false negatives are preferred; the two-phase thresholding with a higher value can be used for better results.

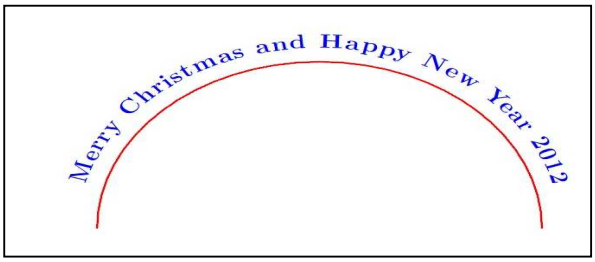
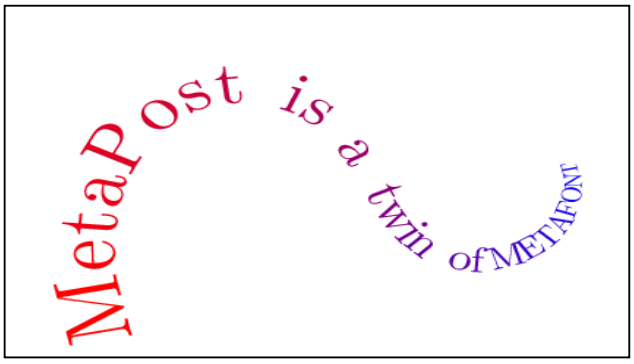


Fig.5(a). Input document

Fig.5(b). Output document



Fig,5(c). Input document

Fig.5(d). Output document

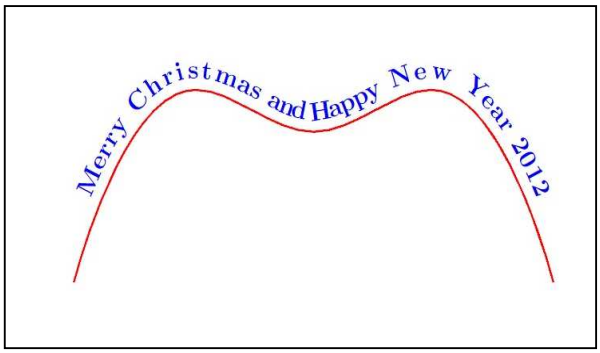


Fig.5(e). Input document

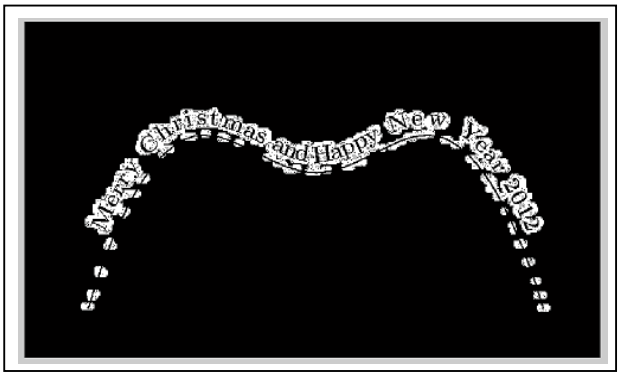
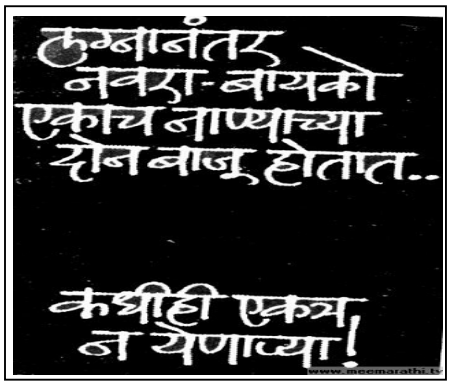


Fig.5(f). Output document



Fig,5(g). Input document

Fig.5(h). Output document

Fig,5(i). Input document

Fig.5(j). Output document

V. CONCLUSIONS AND FUTURE SCOPE

A robust method for text line segmentation from the Multioriented and curved text documents. This method is script independent as well as robust against font, size and color of the text in the document.

Our method cannot extract text from documents properly with uneven lighting, shadowing, low contrast or low luminance present documents.

There is need to find the method for text line segmentation for overlapped text lines or documents, low contrast and low luminance documents. Documents with uneven aliasing, variation in alignment and complexity of background need to be segmented perfectly using different methods or algorithms.

REFERENCES

[1] Neha Gupta, V .K. Banga , "Image Segmentation for Text Extraction," 2nd International Conference on Electrical, Electronics and Civil Engineering (ICEECE'2012) Singapore April 28-29, 2012.

[2] Chung-Wei Liang and Po-Yueh Chen, "DWT Based Text Localization," International Journal of Applied Science and Engineering, pp.105-116, 2004.

[3] S.Audithan, RM. Chandrasekaran," Document Text Extraction from Document Images Using Haar Discrete Wavelet Transform," European Journal of Scientific Research ISSN 1450-216X, Vol.36, 2009.

[4] Keshava Prasanna, Ramakhanth Kumar P, Thungamani.M, Manohar Koli, "Kannada Text Extraction From Images And Videos For Vision Impaired Persons," International Journal of Advances in Engineering & Technology, Nov 2011.

[5] Roshanak Farhoodi and Shohreh kasaei, "Text Segmentation From Images With Textured and colored Background," 13 Iranian conference on Electrical Engineering .

[6] C. Strouthopoulos, N. Papamarkos, A. E. Atsalakis, "Text extraction in complex color documents, " The Journal of Pattern Recognition 35, 1743–1758, 2002.

[7] Chitrakala Gopalan and D. Manjula, "Contourlet Based Approach for Text Identification and Extraction from Heterogeneous Textual Images," International Journal of Electrical and ElectronicsEngineering 2(8), pp. 491-500, 2008.

[8] U. Pal and Partha Pratim Roy, "Multioriented and Curved Text Lines Extraction From Indian Documents," IEEE Transactions on systems, man, and cybernetics—part b: cybernetics, vol. 34, no. 4, august 2004.

[9] Yi Li, Yefeng Zheng, David Doermann, and Stefan Jaeger "ScriptIndependent Text Line Segmentation in Freestyle Handwritten Documents," IEEE transactions on pattern analysis and machine intelligence, vol. 30, no. 8, august 2008.

[10] Mohieddin Moradi, Saeed Mozaffari, and Ali Asghar Orouji, "Farsi/Arabic Text Extraction from Video Images by Corner Detection," 6th, IEEE,Iranian conference on Machine Vision and image processing ,Isfahan, iran,2010.

[11] H. Goto and H. Aso, "Extracting curved lines using local linearity of the text line," Int. J. Doc. Anal. Recognit., vol. 2, pp. 111–118, 1999.

[12] L. A. Fletcher and R. Kasturi, "A robust algorithm for text string Separation from mined text/graphics images," IEEE Trans. Pattern Anal. Machine Intell., vol. 10, pp. 910–918, Nov. 1988.

[13] "Detection of curved text path based on the Fuzzy Curve Tracing (FCT) algorithm," in Proc. 6th Int. Conf. Document Analysis Recognition, 2001, pp. 266–269.

[14] F. Hones and J. Litcher, "Layout extraction of mixed mode documents," Mach. Vis. Applicat., vol. 7, pp. 237–246, 1994.

[15] K. Kise, W. Iwata, and K. Matsumoto, "A computational geometric approach to text-line extraction from binary document images," in Proc. IAPR Workshop Document Analysis Systems, 1998, pp. 364–375.

