Text Localization and Extraction from Complex Color Images

S Sabari Raju, P B Pati* and A G Ramakrishnan
Department of Electrical Engineering
Indian Institute of Science
Bangalore, INDIA – 560 012.

Abstract. Availability of mobile and hand-held imaging devices, such as, cell phones, PDA's, still and video cameras have resulted in new applications, where the text present in the acquired images is extracted and interpreted for various purposes. In this paper, we present a new algorithm for automatic detection of text in color images. Proposed system involves Gabor function based multi-channel filtering on the intensity component of the image along with Graph-Theoretical clustering applied on the color space of the same image, there-by utilizing the advantages of texture analysis as well as those of connected component for text detection. Our approach performs well on images with complex background.

1 Introduction

Embedded text in images and videos provides useful information for automated annotation and indexing. Such text eventually helps in mining the relevant images from the database [1]. An efficient scheme for detection, localization and extraction of textual information from images is needed for such a task. Images may have text of various styles and sizes in simple or complex background. Document analysis softwares, generally, assume high scanning resolution, high quality document images with fairly simple layout structure. These assumptions do not hold for camera captured scenic or document images. Performance of any Optical Character Recognizer (OCR) greatly depends on such a text localization task.

Numerous approaches on text localization have been reported in the literature. Majorly, color and texture are the information being employed for this task. Messelodi *et. al.* [2] have extracted connected components to characterize text objects, based on their size information, in book cover color images. They utilize the heuristics which depend both on the geometrical features of a single component as well as its geometrical and spatial relationship with other components. Zhong *et. al.* [3] segment the image into connected components with uniform color. They use several heuristics on size, alignment and proximity to select the components as likely text characters. In the gray-scale version of the

* Corresponding Author: pati@ee.iisc.ernet.in

obtained image, the regions with higher local variance are termed as text regions. The algorithm is reported to work fine for (a) CD and book cover images, and (b)traffic scene videos. Jain and Yu [4] extract a set of images by analyzing the color space of the input image. They employ connected component (CC) analysis on each of the derived images to locate possible text regions. Finally, they merge the information so obtained to locate text regions in the original image. Strouthpoulos *et. al.* [5] have proposed a technique to determine the optimal number of unique colors present in the input image. In the first step, an unsupervised neural network clusters the color regions. In the subsequent step, a tree-search procedure, using *split-and-merge* conditions decides whether color classes must be split or merged. They use a page layout analysis technique, on each of the obtained optimal color images. Finally, they add the information obtained from each of the optimal color images to extract the text region. Smith [6] uses vertical edge information for localizing caption text in images. Jung [7] has used a neural network based filtering technique to classify the pixels of input image as belonging to text or non-text regions.

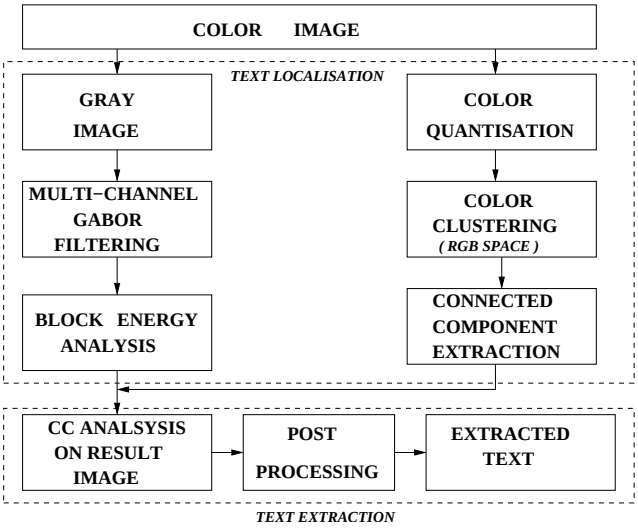


Fig. 1. Overview of the proposed approach.

2 System Description

The proposed text extraction scheme (refer Fig. 1) is demonstrated with texture based method and color clustering. It can be assumed that text regions in an image contain a lot of abrupt changes in the gray values, in various directions, making it rich in edge information [8]. So an ideal feature to discriminate between text and non-text areas should invariably involve directional frequency.

Gabor function based filters are well known to accomplish this task of directional frequency selectivity. So, in the proposed algorithm, Gabor filter based multi-channel filtering is adapted followed by Block Energy Analysis (BEA), which is employed on the energy image, *i.e.*, output of the Gabor filter. Parallelly, the color image is quantized and color components in the image are analyzed using Graph-Theoretical cluster technique. The results of the Gabor based multi-channel filter and color component analysis are merged, and CC analysis is done based on geometrical and statistical information of the individual components. Components which are present inside the bounding box of another component are removed at the post-processing stage.

2.1 Gabor Filter Bank

In a given image, consisting of text and non-text regions, the text regions are quite rich in high frequency components. The technique involving multi-channel filtering with Gabor function based filters, for page layout analysis, is developed due to convergence of biological [9, 10] and machine learning approaches. Such a method is meant to detect text regardless of the type of script, size of font or the layout the text is embedded in, and is more robust than other kinds of feature detection models. Besides, it has a low sensitivity to various kinds of noise.

A bank of Gabor filters are chosen for extraction of the above mentioned features. A Gabor function is a Gaussian modulated by a complex sinusoid.

$$h(x,y) \;=\; g(x',y')\; exp[j2\pi\; Ux] \tag{1}$$

where x' and y' are the rotated components of the x and y co-ordinates, in the rectangular co-ordinate system, rotated by angle θ. U is the radial frequency in cycles/image width.

$$g(x,y) = \frac{1}{(2\pi\sigma_x\sigma_y)} exp\left(-\frac{1}{2}\left[\left(\frac{x}{\sigma_x}\right)^2 + \left(\frac{y}{\sigma_y}\right)^2\right]\right) \tag{2}$$

σ_x and σ_y explain the spatial spread and the bandwidth of the filter function $h(x,y)$. If B_r is the radial frequency bandwidth in octaves and B_θ is the angular bandwidth in degrees, then,

$$\sigma_x \;=\; \frac{\sqrt{2}}{2\pi U}\frac{2^{B_r}+1}{2^{B_r}-1}\;, \qquad \sigma_y \;=\; \frac{\sqrt{2}}{2\pi U \tan\left(\frac{B_\theta}{2}\right)} \tag{3}$$

The power of discrimination of the different filters are dependent on the values of B_r and B_θ.

Any combination of B_r, B_θ and U involves two filters, corresponding to the sine and the cosine functions respectively, in the exponential term in Eqn. 1. Gabor filters of $B_r = 1$ octave and $B_\theta = 45^o$ at four different orientations ($\theta = 0^o,\ 45^o,\ 90^o,\ 135^o$) have been used for the reported work. Three different radial frequency, U, values ($U = \ 0.2,\ 0.3\ and\ 0.5$) have been chosen as they have been observed to be working well for all kinds of images used in our work.

2.2 Block Energy Analysis (BEA)

A *space-frequency* filter bank, using Gabor filters, has been designed and implemented for separating the text and non-text areas in a gray level document image. The filter response of a complex filter (refer Eqn. 1) at any pixel (i, j) is:

$$E(u_l,\ \theta_k) \ = \ \sqrt{e(u_l, \theta_k)^2 + o(u_l, \theta_k)^2} \tag{4}$$

where $e(u_l, \theta_k)$ and $o(u_l, \theta_k)$ are the outputs of the cosine(even) and sine(odd) filters, with radial frequency u_l and angle θ_k, at that pixel location, respectively. The total local energy is the sum total of the outputs of all the complex filters at that given pixel.

$$E_T \ = \ \sum_{l=1}^{3} \sum_{k=1}^{4} E(u_l,\ \theta_k) \tag{5}$$

We low pass filter the magnitude response image of the Gabor filter bank (the E_T image) with a Gaussian filter. A 11×11 Gaussian mask ($\sigma_x = \sigma_y = 3$) is employed for this purpose. It is seen that this removes artifacts and irregularities present and, thereby, helps in text detection. We evaluate block energy at each pixel by considering blocks of 15×15 size. An average of the block energies is evaluated across all the blocks in the image. Twice this average energy (decided heuristically) is the threshold for block wise text and non-text separation, *i.e.*, if the block energy is found to be above the threshold value, thus calculated, the block is considered as text area.

2.3 Color Component Generation and Analysis

The methodology for the color component generation is as follows:

1 Quantize the color space into prototype colors. The prototypes are found by masking the two least significant bit.
2 Compute the 3-D histogram (RGB color space with 256 pixel level) of the color image.
3 Construct lists of pixels representing each histogram bin.
4 For every unique color in the histogram, link to the bin which has maximum vote in the neighborhood of 20, forming non-overlapping unimodal clusters in tree structure. (The self-referencing bins are the local maximum and representative of the cluster, C_i, $i = 1,\ 2,\ \cdots,\ N$.)
5 Generate connected components from each cluster to construct binary images
6 Connected Component Analysis:
    FOR each cluster
        N_i is the no. of ON pixels in i_{th} cluster.
        $N_m \ = \ Max\ \{N_i\};\ i = 1,\ 2,\ \cdots,\ N.$
        If ON pixel in the cluster, N_i, is greater than 5% of N_m.
            FOR each connected component in the cluster:
            Select text components based on geometrical and statistical information of the components.

END FOR each
END IF each
END FOR each

7 At the end of the CC analysis, final image is constructed by merging all components obtained from different color planes.

3 Experimental Results and Discussion

Our image database consists of 150 color images. These images are (i) scanned with 300 dpi and stored at 24-bit color representation, (ii) captured with **Minolta Dimage 2330 Zoom** still camera, (iii) captured with the camera present in the **Sony Ericsson T610** mobile phone, (iv) frames extracted from videos and (v) downloaded from WWW sources. The proposed algorithm performs consistently well on a majority of these images.

The algorithm is evaluated by counting the number of detected characters. The following *rates* are formulated as a quantitative measure for the effectiveness of the algorithm.

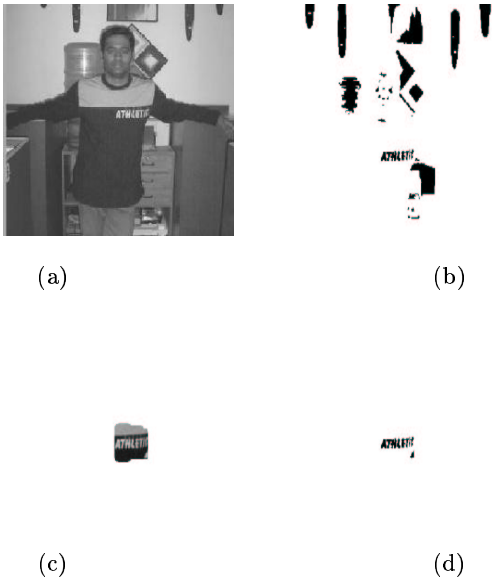
$$Precision\ Rate\ =\ \frac{No.\ of\ correctly\ detected\ text\ characters}{No.\ of\ detected\ characters}\ X\ 100 \quad (6)$$

$$Recall\ Rate\ =\ \frac{No.\ of\ correctly\ detected\ text\ characters}{No.\ of\ characters\ present\ in\ original\ Image}\ X\ 100 \quad (7)$$

The accuracy of detection is evaluated on 70 images from the database. 91.7% of precision rate is attained in an average while the average recall rate achieved is 89.2%.

Fig. 2 illustrates the comparison between general color clustering technique with CCA and the proposed algorithm. A camera-captured image is shown in Fig. 2(a) which has very less text in the image. Fig. 2(b) shows the text region using color clustering and CC analysis and Fig. 2(c) shows the output using the Gabor filters. Fig. 2(d) shows the output of the proposed system where regions of text are well detected in-spite of the dominance of natural background.

A natural image with synthetic text is shown in Fig. 3(a). This image has 40098 unique colors. The interesting point of this example is the gradual change of the background colors. The application of the proposed text extraction technique, gives results shown in Fig. 3(d). Test image 3(b) has background of two principal color tones. In addition, this document has white as text and background regions. Complexity in the connected component analysis here is the background of the black text region (bounding box of the character) has white color, which also satisfies the characteristics of character. The text areas are correctly obtained and are shown in figure 3(e). Fig. 3(c) shows a color image with complex background. The number of unique colors in this image is 88399.



(a) (b)

(c) (d)

Fig. 2. Text localization Results (a) Original Camera Image (b) Result only with Color Clustering and standard connected component analysis (c) Output of the Gabor filter (d) Result of proposed scheme.

It can be observed that the majority of the text areas are correctly obtained and this is shown in Fig. 3(f).

Table 1 presents the time taken by each module of the algorithm for a typical camera captured image (520×640) and for a frame extracted from a video sequence of size 240×320. It may be noted that the Gabor filtering part of the algorithm is computationally expensive.

Table 1. Processing Time for Video Frame (240×320) and Camera Image (520×640) in seconds.

Operation	**Video Frame (secs)**	**Camera Image (secs)**
Gabor Filtering	3.0	7.0
Computing 3-D histogram	0.03	0.035
Constructing and Linking Tree	0.37	0.45
CC analysis in one cluster	0.07	0.08

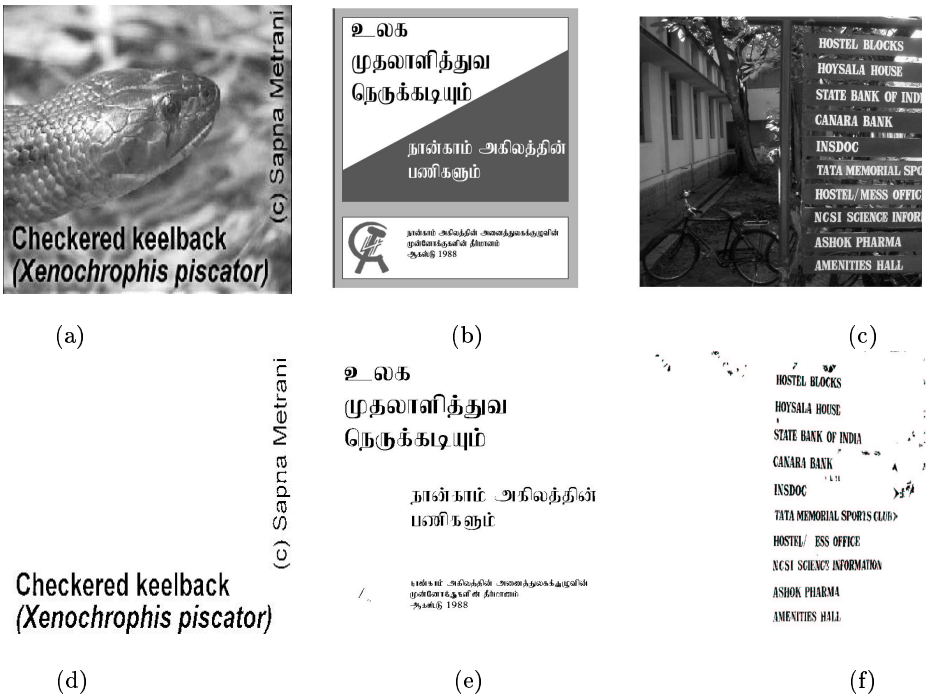


Fig. 3. Experimental Results (a) WWW image (b) scanned book cover image (c) Camera Image (d)-(f) Extracted text regions.

4 Conclusion

In this paper, an approach to automatically localize and extract text, on camera captured (still and mobile) as well as on scanned document images, is proposed and implemented. The results obtained thus far are encouraging.

The algorithm is able to detect and localize text in document images as well as in scenic images with complex background. It has the ability to deal with high resolution scanned images and camera captured images with equal efficiency. Not only is the proposed algorithm script invariant, it also is invariant to skew present in the document images. The disadvantage of our algorithm is the computational cost. The running of the algorithm on the mobile images shown in figure 2 takes about $3.5s$ and camera captured images shown in Fig. 3(f) takes $7.82s$ on a PC with Pentium-IV processor at 2.2 GHz with 512 MB of RAM. When text extraction performance is the criteria, our scheme performs consistently well.

Integrating an OCR which can perform well on low-resolution images with the proposed text extraction method is proposed as a scope for future investigations. The proposed approach could also be applied to handwritten documents.

References

1. Antani, S., Kasturi, R., Jain, R.: A survey on the use of pattern recognition methods for abstraction, indexing and retrieval of images and video. Pattern Recognition **35** (2002) 945–965
2. Messelodi, S., Modena, C.M.: Automatic identification and skew estimation of text lines in real scene images. Pattern Recognition **32** (1999) 791–810
3. Zhong, Y., Karu, K., Jain, A.K.: Locating text in complex color images. Pattern Recognition **28** (1995) 1523–1535
4. Jain, A.K., Yu, B.: Automatic text location in images and video frames. Pattern Recognition **31** (1998) 2055–2076
5. C.Strouthpoulos, N.Papamarkos, Atsalakis, A.E.: Text extraction in complex color document. Pattern Recognition **35** (2002) 1743–1758
6. Smith, M.A., Kanade, T.: Video skimming for quick browsing based on audio and image characterization. Technical report, Technical Report CMU-CS-95-186, Carnegie Mellon University (1995)
7. Jung, K.: Neural network-based text location in color images. Pattern Recognition Letters **22** (2001) 1503–1515
8. S, S.R., Pati, P.B., Ramakrishnan, A.G.: Gabor filter based block energy analysis for text extraction from digital document images. In: Intl. Workshop on Document Image Analysis for Libraries. (2004)
9. Porat, M., Zeevi, Y.Y.: The generalized gabor scheme of image representation in biological and machine vision. IEEE Trans. on PAMI **10** (1988) 452–467
10. Morrone, M.C., Burr, D.C.: Feature detection in human vision: a phase dependent energy model. Proceedings of the Royal Society of London(B) **235** (1988) 221–245