An Efficient Method for Text Extraction from Colored Images

Shilpi Rani
CS & IT deptt.
IFTM University
Moradabad

Rakesh Kumar Yadav
CS & IT deptt
IFTM University
Moradabad

ABSTRACT
Text Extraction from image is concerned with extracting the relevant text data from a collection of images. Due to rapid development of digital technology digitization of all these images with text is necessary. Lot of resources such as newspapers, books, journals records business card, magazines, advertisements slides and films, scanned document, book covers etc are converted to images and are available in electronic medium. Text extraction and recognition from these images present many challenging research issues .The proposed system extracts text from colored images using edge based technique and morphological operations. This system is implemented using matlabR2013a and is better than existing system because it can also extract the text from complex colored images This paper explains the necessary steps required to extract text from colored images

General Terms
Color Transformation, Edge Detection, Text Localization, Text Extraction, Text region, non text region.

Keywords
Binarization, Blobs, image, image processing, Optical character recognition, Thresholding,

1. INTRODUCTION
In today's era, the information libraries that originally contained pure text are becoming increasingly enriched by multimedia components such as images, videos and audio clips. An automatic means is required to efficiently index and retrieve multimedia components from all these multimedia resources. They would be a valuable source of high level semantics if the text occurrences in images could be detected, segmented, and recognized automatically. Color images that integrate text and graphics communicate in an immediate and effective manner and are widely used [1]. However, such images are often a complex mixture of shapes and colors arranged in unpredictable ways, which make it difficult to automatically extract or separate the text from the rest of the color image. Images can be classified into document image, caption text image and scene text images.
A Document image (fig1 &2) Documents with text embedded in complex colored and textured backgrounds are increasingly common today, for example in magazines, newspapers, magazines and web pages. From these documents text detection is a challenging problem. The approaches developed, such as binarization by adaptive thresholding, for ordinary documents are not generally applicable, because with these technique it seems to be difficult to find an optimal threshold or thresholds to preserve meaningful information and to eliminate unnecessary one Document image contains only text and some graphics. The document images may contain unlimited number of fonts, style, alignment, size, shapes, colors, etc. Extraction of text in documents with text on complex color background is difficult due to complexity of the background and mix up of colors of fore-ground text with colors of background.
Caption text is also known as Overlay text or Cut line text. Caption text (Fig 3) is artificially superimposed on the video/image at the time of editing and it usually describes or identifies the subject of the image/video content [2]. Scene text (Fig 4) appears within the scene which is then captured by the recording device i.e. text which is present in the scene when the image or video is shot. Scene texts occurs naturally as a part of the scene and contain important semantic information such as advertisements that include artistic fonts, names of streets, institutes, shops, road signs, traffic information, board signs, nameplates, food containers, cloth, street signs, bill boards, banners, and text on vehicle etc.

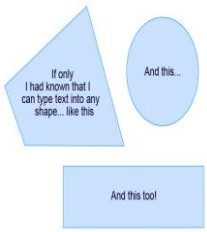

Fig 1: document text image

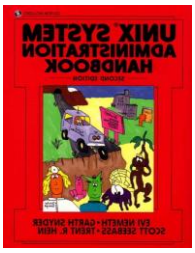

Fig 2: colored text image

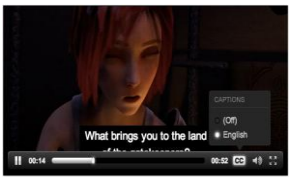

Fig 3: caption text image

Fig 4: scene text image

1.1 Text Extraction System
The TIE problem can be divided into the following sub-problems: (i) detection, (ii) localization, (iii) tracking, (iv) extraction and enhancement, and (v) recognition (OCR) [3].

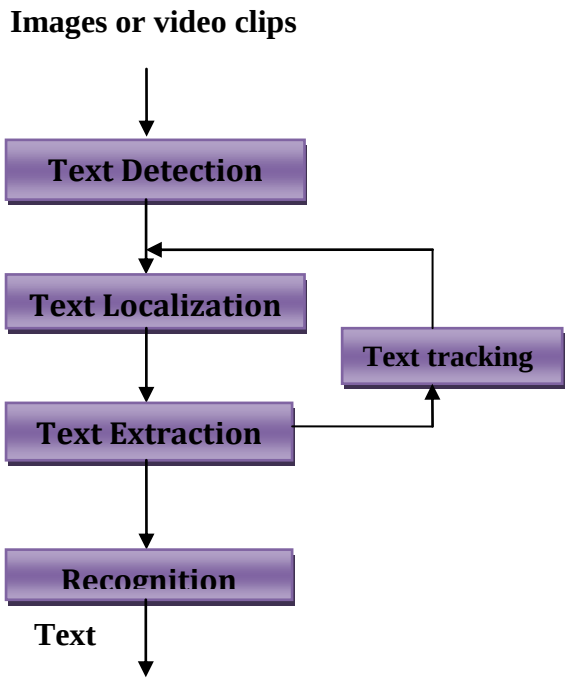

Fig 5: Text Extraction System

An important step for document image analysis and recognition is text binarization. The useful content of the image that conveys some information is text .The important characteristics of text are edge motion, size and color. The process of converting an image (up to 256 gray levels) to black and white images (0 or 1) is binarization. The conversion of a gray-scale or colored image into a textual binary image is referred as document image text binarization.

A binarization process precedes the analysis

and recognition procedures in most document processing systems. The use of two-level information in the form of 0 and 1 greatly diminishes the complexity and computational load of analysis algorithms. To achieve robust binarization is difficult, because any error introduced in this stage will affect the subsequent processing steps. The texts are discriminated from the background in document images applying binarization methods. In images graphics also serves like a source of non text. In colored document Documents text binarization is not merely conversion to black and white but it also includes removal of non textual elements like images, graphics, background noise, shadows, table boundaries etc. Due to the advancement of color printing technology, color document images are employed increasingly. Optical character recognition (OCR) techniques are used for document digitization and recognition [4]. For most OCR engines, character features are extracted and trained from binary character images. It is relatively difficult to obtain satisfactory binarization images from various kinds of document images. Moreover, it is difficult to extract text from these color document images.

The ultimate goal of document binarization is to convert a given grayscale or color document image into a bi-level representation. The underlying objective is to separate objects, like characters, from the background with the assumption that gray levels of pixels belonging to the two classes are substantially different. The quality of the binarization is crucial for document recognition because most of the algorithms used during analysis (page orientation, layout analysis, character recognition, etc.) assume a black and white image and rely on the output of the binarizer [5].

1.2 Need for Text Extraction

The rapid growth of personal and professional multimedia databases led to the need for text extraction from colored documents. Document images no longer comprise of only textual content, varying non textual content like images, graphs, tables are also embedded in them. The research community devoted lots of effort to content-based image indexing, but bridging the semantic gap is difficult. The low level descriptors used for indexing (e.g.: interest points, texture descriptors) are not enough for an efficient manipulation of big and generic image databases. Document processing is an active research area with significant applications

1.3 Applications

Text extraction from images have many useful applications in detection of vehicle license plate ,analysis of article with tables, maps, document analysis charts, diagrams etc. keyword based image search, identification of parts in industrial automation , page segmentation ,content based retrieval, name plates, object identification, street signs, text based video indexing, video content analysis, document retrieving, address block.

2. CHALLENGES

Several problems that arise during text extraction from colored images are:

- Document is composed of multiple colors
- Background and foreground are of same color
- Background having multiple colors
- Rounded text, multioriented text, curved text etc.
- Text in multiple fonts and sizes
- Text in both upper and lower cases
- Text in multiple languages

Various approaches have been proposed for text extraction from images. But still there is no single approach that can deal with all these sorts of problem therefore it's an growing research area .Many factors degrade the performance of textual information extraction techniques. For scanned document images, salt and pepper noise, touching and broken characters, and skew have long been the processing obstacles[6]. For camera-based images, low-resolution, blur, warping, as well as perspective distortion are the major challenges. Among these degradation, interested one is geometric deformations, i.e. skew and perspective distortion. Skew may be generated in a scanned document image if the edge of the paper is not aligned correctly with the scanner during scanning [7]. Perspective deformation of a camera-based document images caused by the fact that the image plane in the camera is not parallel to the document plane, and manifests as severe skew, unpredictable orientation, non-parallel text-lines.

3. PROPOSED APPROACH

The proposed approach is mainly divided into three steps.
First is Preprocessing, second is Text detection and localization, and third is finally Text extraction.

Preprocessing mainly deals with removal of noise and correcting errors from images. It is mainly done in the form of color transformation, noise removal.

Text detection and localization mainly deals with detecting connected components and drawing bounding boxes around the text.

Text Extraction means extracting only those boxes that contain text .This enables us to separate text region from non text region

3.1 Gray Scale Conversion

Gray scale conversion is mainly done using standard Grey (lum) =0.3R+.59G+.11B.
Where R, G and B corresponds to red, blue and green color respectively with values ranging from 0 to 255

Fig 6: Input Image

Fig 7: Color transformation

3.2 Edge Detection

Edge detection process is responsible for detecting possible edges and boundaries in an image. This process is based on the assumption that text has vertical or horizontal edges and the best approach to start detecting any text part in an image is to check for edges. Since an image contains information in almost every part of it, so the initial task is to filter out those parts where the system is sure that no text exists. This process allows the system to focus on selected parts of the image rather than on the complete image. However it does not process the image for any text detection, but just generate a normal edge image against each frame. This process detects those pixels which have certain difference in their value compared to their surrounding pixels. Canny edge detection algorithm is used to detect edges in images.

Fig 8: Edge detected image

3.3 Morphological Operations

In this approach boundary extraction opening and closing operators are used for removing unwanted and noisy components and efficient extraction of text.

Fig 9: Image after morphological operations

3.4 Text Detection and Localization

Text detection is done here using connected component based approach. Connected components are detected and labeled .The blobs are also created around these components. Text Localization is done using Horizontal profiling, Checking for text alignment; horizontal or vertical and finally applying geometric and shape regularity features.

Fig 10: Image with blobs

3.5 Text Extraction

Text extraction is done mainly by retaining only those blobs which contain text.

Fig 11: Text extraction

Table 1: Result analysis

Type of Image	No of Images	Text Detected completely	Text detected along With non text	Text not detected
Colored text images	60	60	0	0
Cover images	80	38	30	12
Images with both text and non text (English)	95	85	10	0

4 CONCLUSIONS AND FUTURE WORK

Several methods exist for extraction of text in images. The proposed approach used edge based and morphology based techniques to improve accuracy. Different attributes related to text in an image such as of size, font, style, orientation, alignment, contrast, color, intensity, connected-components, edges etc are used to classify text regions from their background or other regions within the image.

The future work may focus on text extraction directly from colored images and extraction of multioriented text also.

5 REFERENCES

[1] Nirmala Shivananda and P. Nagabhushan, "Separation of Foreground Text from Complex Background in Color Document Images," IEEE Transactions on Image Processing, vol. 10, pp. 306-309, 2009

[2] Paraag Agrawal, Rohit Varma "Text Extraction from Images" IJCSET |April 2012| Vol 2, Issue 4, 1083-1087

[3] Kohei Arai, Herman Tolle (2011)," Text Extraction from TV Commercial Using Blob Extraction Method", International Journal of Research and Reviews In Computer Science Vol.2, No.

[4] Shivani Saluja, Tushar Patnaik, Tanvi Jain "Text Extraction and Non Text Removal from Colored Images ", International Journal of Computer Applications (0975 – 8887) Volume 44– No.22, April 2012

[5] Yen Len chin "Automatic Text Extraction, Removal and Inpainting of complex document images" International Journal of Innovative Computing, Information and control volume 8, Number 1(A), January 2012

[6] Md. Shorif Uddin, Tanzila Rahman, Umme Sayma Busra and Madeena Sultan "Automated Extraction of Text from Images using Morphology Based Approach" (IJEI) 14th august 2012 vol.1,No.1,2012

[7] Sachin Grover,Kushal Arora,Suman K.Mitra" Text Extraction from Document Images using Edge Information" IEEE India Council conference 20 December 2009.

[8] Huang, Huadong Ma, He Zhang, "A New Video Text Extraction Approach" IEEE International Conference on Multimedia and Expo, 2009.ICME 20092011.

[9] Chen D, H. Bourlard, 2001. And J. -P. Thiran, "Text Identification in Complex Background using SVM", Proc. of IEEE Conference on Computer Vision and Pattern Recognition, Vol. 2, pp. 621-626.