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].