Chapter 1

Introduction
INTRODUCTION

1.1 Getting started

Nowadays, with the advancement of digital technology more and more databases are multimedia in nature. The databases usually contain images and videos in addition to the textual information to understand those images. Textual information in an image is an important source because it contains very useful high level semantic information [K96] compared to visual and audio information. For example texts embedded in news videos usually simplify the content of the news reports. In addition, Optical Character Reader (OCR) is more robust than automatic speech recognition and visual analysis techniques. So an automatic extraction of text from images can be used to fully understand the images and videos without human supervision. Contents in image can be perceptual, such as color, shapes, textures, size, etc., or semantic, such as text, objects or events and its relationships. The perceptual contents are easier to analyse automatically while the semantic contents are easier to handle linguistically. Since 1990s, with the rapid growth of available multimedia documents and an increasing demand for information indexing and retrieval, a lot of effort has been taken on text extraction in images and videos. Manual annotation of videos and images is extremely time consuming, expensive, and unscalable in the features of
ever growing image databases. Therefore, an automatic extraction of text from image descriptions is desirable in order to annotate and search large image databases often superimposed on the frames in textual form.

1.2 Background and Applications of Text Extraction

Text regions in an image may carry useful information about the visual content. Accordingly, extraction of text has become an important tool in a vast number of applications [KIK04] including indexing and retrieval, document analysis, vehicle license plate extraction, object-oriented data compression and character extraction [SK95] [NN97] [CDP99] [TTKS98] [YI99]. In the following paragraphs, some of these applications are described.

*Text searches in Images*: Text extraction would enable better searching by extracting the content of an image. Watanabes [YYBT98] translation camera can detect text in a scene image and convert Japanese text into English after the stage character recognition.

*Content based Indexing*: For the purpose of archiving and indexing images and videos, the content of the image is required in the digital format. Knowledge about the text content of images can help
in the building of an intelligent system which archives and indexes the images. This involves automatic text-based video structuring methods using caption information [TTKS98] [BC95].

**Reading foreign language text**: One of the common problems faced by a person in foreign is that of communication, understanding road signs, License Plate, signboards etc [BKM97]. Text extraction method uses to alleviate such problems by reading the text information from the image scenes which are captured by a camera. Kim et al. [SDYG02] used gradient method to extract text region information from car license plate information under complex image conditions.

**Archiving documents**: Collection of paper documents in offices or other printed material like newspapers and magazines can be electronically transformed for more efficient storage and instant delivery to office or home computers. Readers may refer to papers on document segmentation [KY96] [YHY96] for more examples of page layout analysis.

**Content-based document coding**: The MPEG standard image supports object-based encoding. When text regions are segmented and extracted from other regions in an image, this can provide high compression rates and higher image quality. Feng et al. [GHC01] and Cheng et al. [HAP97] apply adaptive dithering after segmenting a
document into different classes. As a result, they can achieve a better quality rendering of documents containing text and pictures.

**Industrial automation**: Automatic inspection of parts done in some applications has shown to be more reliable than human visual inspection e.g. given the image of a printed circuit board, a computer vision system can determine proper placement of components [KSKH95].

### 1.3 Text in Images

Text in images can be classified into caption text (Figure 1.1(a)) or artificial text or superimposed text which is artificially overlaid on the image and scene text (Figure 1.1(b)) or graphics text which occur naturally in an image [SDA + 00][S01]. In comparison to the caption text, scene text in an image or a video frame can have a different direction and might be distorted by the perspective projection. Additionally, it is often affected by variations in scene and camera parameters such as illumination, varying light, transformation, focus, motion, etc. Therefore it is more difficult to detect the text region in an image and very little work has been done in this area. An added type of image such as scanned document images deals with page layout analysis [CNE02]. Readers may refer to papers on document segmentation [KY96] [YHY96] for more examples of
document images (Figure 1.2). The images acquired from CD covers, scanning book covers or other multi-colored documents have similar characteristics as the document images. Text extraction can also be adopted and extended for video images [RF96] [UDS + 99] [KHW00] and it is difficult due to many undesirable properties of video documents such as size, position, orientation, low resolution, low contrast, text color and cluttered backgrounds. The scene text in a video image occurs naturally in the 3-D scene as it is naturally recorded and captured by video camera. The caption text in a video image comprises of 2-D strings that are composited on to the video frame during the editing stage of production.

Figure 1.1: (a) Caption Images b) Scene Text Images

Figure 1.2: Document Images
1.4 Characteristics of Text in Images

Text present in images usually have different appearance changes like font, size, style, orientation, alignment, texture, color, low image contrast, and a complex background. All these changes will make the problem of automatic text extraction difficult and complicated. They can exhibit many geometric variations with respect to the following properties [DSR01] [DJK00] [ZZJ00] [SUD + 99] [USR98].

Size: Even though the text size can differ a lot, assumptions can be made depending on the application domain.

Alignment: The characters in the caption text appear in images as in the form of clusters and usually lie horizontally and also they can come out as non-planar texts as a result of special effects and projections. This is not related to the scene text in images which can have various perspective distortions and illuminations. It can also be aligned in any direction and can have more geometric variation properties.

Inter-character distance: Characters present in the text area have a uniform spacing between them.

Color: The characters in a text region tend to have the same or like colors. This color property makes it possible to detect text using a connected component-based technique. Most of the study reported till date has intense on finding text region of a particular color (monochrome). However, video frames and other complex color images
can contain text region with various colors (polychrome) for effective visualization, i.e., multi colors within one word.

**Motion:** The same texts are usually present in consecutive frames in a video with or without movement. This property can be used in text tracking and enhancement. The caption text usually moves in the same way: vertically or horizontally. The scene text can have different arbitrary motions due to camera or object movement.

**Edge:** Scene and caption texts of connected components can be easily understood by finding the edges in the boundaries of the background and the appearing text region in the image.

**Compression:** Many images or videos are recorded, transferred, and processed in a compressed file format. Thus, a faster text extraction information system can be achieved if one can extract the text region without decompression.

### 1.5 Text Information Extraction (TIE)

A TIE process receives an input in the form of a still image or a sequence of images (video). The images can be in compressed or uncompressed, color or grey scale, and the text in the images may or may not move. The TIE system can be divided into the following stages [KIK04]: i) detection, ii) localization, iii) tracking, iv) extraction and enhancement, and v) recognition (OCR)
**Text detection** is the determination of the occurrence of the text in a given image or video frames.

**Text localization** is the method of determining the position of the text in the image and generates the bounding boxes around the text.

**Text tracking** is performed to ease the processing time for text localization and to keep the integrity of its position across the adjacent frames. Even though the precise location of the text in an image can be indicated by bounding boxes, the text still has to be segmented from the background to facilitate its recognition. Thus extracted text image has to be changed to a binary image and enhanced before it is fed into an OCR engine.

**Text extraction** is the step where the text components are segmented from the complex background of the image.

**Text Enhancement** of the extracted text region components is essential because the text region in an image has low resolution and is prone to noise and can be transformed into plain text using OCR technology.

In the above stages, text detection and text localization are closely associated and more challenging stages which attract the interest of most researchers. The objective of these two stages is to make accurate bounding boxes of all text objects present in image or video frame and to provide an unique identity to each text. Following these stages the text is extracted and filters out the background pixels in the text area.
1.6 Thesis Objectives

The text extraction in images and videos is a significant research branch of text-based image indexing and content-based information retrieval continue to be an area of much interest to researchers. A complete text-image-analysis is needed to enable a text information extraction system to be used for any type of image and video, including both scanned document images and real scene images through a video camera. In addition, the text enhancement is also needed for low quality images and more adaptability is needed for general cases (e.g., 2D or 3D deformed characters, inverted characters, polychrome characters, and so on). Otherwise the text extraction results are inappropriate for general OCR software. In spite of many difficulties in using TIE systems in real world applications, the significance and usefulness of this field continues to attract much attention. In this present research, an efficient combined method has been proposed for automatic text content extraction from different images which is independent of the geometric variations, orientation of the text, low contrast images, complex background of the image as well as font size of the text.

1.7 Thesis Organization

The rest of the thesis is organized as follows: With rapid developments in text extraction process in images and videos, several research works that have been summarized and discussed in the literature and their
recent progress based on TIE techniques are reviewed in Chapter 2. Chapter 3 describes the fundamental methodologies of text extraction process which are used in the current research work. Subsequently, three set of experiments have been carried out in this research work. The first set of experiments is done in the scene and caption text images discussed in Chapter 4. Chapter 5 shows the second set of experiments in noisy video images. The third phase of the experiments in documents images is discussed in Chapter 6 using an effective combined technique. Chapter 7 presents the experimental results and comparisons showing that the proposed combined technique outperforms in different noises using different filters in terms of metrics and gives a good detection rate, low fragmentation and low error rate.