THE INTERNATIONAL JOURNAL OF SCIENCE & TECHNOLEDGE

A Survey on Text Extraction from Images

Garima Rana
Deartment of Computer Engineering
Vishwakarma Institute of Information Technology, University of Pune, India
Chaitali Gandhi
Deartment of Computer Engineering
Vishwakarma Institute of Information Technology, University of Pune, India
Poojan Analdas
Deartment of Computer Engineering
Vishwakarma Institute of Information Technology, University of Pune, India

Abstract:
This is survey paper on text extraction from images. A large number of algorithms and methods are proposed to extract text from the given image. Extraction of this information involves text region detection, text localization, tracking, character extraction, enhancement, and recognition of the text from a given image. Variations in text may occur because of differences in size, style, orientation, alignment of text, and low image contrast, composite backgrounds make the problem during extraction of text. The purpose of this paper is to classify and review various text extraction algorithms, discuss working and performance evaluation, and finding a technique for getting maximum accuracy.

Key words*: Text extraction, Text detection, Text localization, Text retrieval, OCR*

1. Introduction

Text in the image contains useful information which helps to acquire the overall idea behind the image. Character extraction from image is important and has many applications.Several methods for text (or character) extraction from natural scenes have been proposed. If we develop a method that extracts and recognizes those texts accurately in real time, then it can be applied to many important applications like document analysis, vehicle license plate extraction, text- based image indexing, etc and many applications have become realities in recent years [1]. Text regions may contain very useful information regarding the image. Before we go for actual text extraction, first we should study the properties of text. Text contains various characteristics like size, motion, color, edge etc.

- Size: differences in text size can make problem, but it can be minimized by assuming specific data during text region detection process.
- Color: intensity of color also affects the quality of text extraction. If all characters of same color then extraction process become more simple and effective.
- Motion: this property of text usually applied to the videos containing text, and this refers to the movement of text in vertical or horizontal direction.
- Edge: edges are reliable feature of text as compared with the other features like color layout or motion

Contents of the paper:

- Section 2 describes various algorithms based on the text extraction
- Section 3 describes Architecture of proposed system is
- Section 4 gives the conclusion .

2. Algorithms Related to Text Extraction

Text extraction process mainly consists of five important phases:

- Text region detection
- Text localization
- Tracking
- Character extraction
- Text recognition.

There are various ways to complete these phases, some of the techniques are:

2.1. Edge Based Text Extraction

Edge based extraction is one of the more efficient method because edge is most reliable feature as compare to others like layout, color, or orientation etc , and this methods focus on the high contrast between background and actual text. The edges of the text boundary are identified and merged, and then several techniques are used to filter out the non-text regions. Some implementations based on edge based technique are as follows:

2.1.1. Algorithm by Xin Zhang et al.

Xin[5] proposed this two phase method: a) Text background removal: For this first transition map model is utilized and to improve the accuracy of text extraction rate of first model second method edge based text detection is used. In this method, two methods are combined, and because it this method called as color-edge combined algorithm. b) Text extraction: in this phase the image is binarized and passed to OCR model for character extraction.

2.1.2. Algorithm by Xiaoqing Liu et al.

Xiaoqing's method [6] consists of three stages:

- Candidate text region detection: . A feature map is binary image where pixel intensity gives possibility of text. In this stage a feature map is generated using three main characteristics of edge viz. strength, density and orientation
- Text region localization: In this stage morphological dialation operator is used. There are two constraints utilized to find non text regions, first for finding very small isolated blocks and second for filter out the block whose width is very small than that's corresponding height.
- Character extraction: Here existing OCR engines were used for character extraction.

This can only deal with printed characters against clean backgrounds and cannot handle characters embedded in shaded, textured or complex backgrounds

2.2 Region Based Text Extraction

A geometrical analysis is done during merging process, to filter out text and non text regions in the image. In region based methods, we consider the properties of colour in text or the variance related to background Many methods are proposed for text extraction which is based on region based text extraction method. .

2.2.1. Algorithm by bunke and Kronenberg[4]

Algorithm for "Identification of Text on Colored Book and Journal Covers",

Color variations are minimised by applying clustering methods in pre-processing step.

- Top down analysis: in this phase the image is split in vertical and horizontal directions alternatively. The output is in rectangular shaped blocks and text containing at least two colors. Depending on this information we reject homogeneous regions means regions having no text.
- Bottom up analysis: it detects homogeneous regions using a region growing method. Beginning with a starting pixel, pixels are merged if they are from to the identical cluster. We know that characters of printed text generally do not touch each other; several regions are detected for a text region. After this the outputs of two methods are combined to distinguish between text and non text regions. After this phase region is binarized using previously gathered information. And this is given as an input to OCR. This method not only limited for book covers, we can use it for other types of images.

3. Architecture for Proposed System

Our method consists of three stages: candidate text region detection, text localization and text extraction. Up till now we had seen various algorithms for text extraction from images but they are proposed for specific applications .There is no general purpose system for text extraction from images and hence the proposed system is very useful.

The three distinguishing characteristics of text embedded in images are edge strength, density and the orientation variance .These characteristics can be used as main components of detecting text

3.1. Image Pre-processing

If the image data is not represented in specific color space ,it is converted to this color space by means of an appropriate transformation.

3.2. Edge Detection

This step focuses the attention to areas where text may occur. Basically the character contours have high contrast to their local neighbours. As a result, all character pixels as well as some non-character pixels which also show high local color contrast are registered in the edge image. We use a simple method for converting the gray-level image into an edge image. We decide a threshold value and compares it with pixel value, according to that the image is converted in binary image.

3.3. Detection of Text Regions

The binary image analyzed in order to locate text areas. In processing, the local maxima are calculated; two thresholds are employed to the local maxima. Finally, the exact coordinates for each of the detected areas are used to create bounding boxes.

3.4. Enhancement and Segmentation of Text Regions

First, geometric properties of the text characters like the height, width, and ratio of width to height are used to discard those regions whose geometric features do not fall into the predefined ranges of values.After that the binary edge image is generated from the edge image, erasing all pixels outside the previously made text boxes and then binarizing it.

4. Conclusion

In this paper we provided various text extraction techniques. Though we have large number of algorithms and methods for text extraction from image but none of them provide adequate output because of deviation in text. The proposed method may give a satisfactory output because it is based on edge detection and it supposed that edge is a more reliable feature of text as compared to others.

5. References

1. F. Chassaing , C. Wolf, J. M. Jolion, and, "Text localization, enhancement and binarization in multimedia documents," in Pattern Recognition, Aug. 2002, vol. 2 of Proceedings. 16th International Conference on, pp. 1037–1040.
2. D. Crandall, R. Kasturi,and S. Antani, Robust Detection of Stylized Text Events in Digital Video, Proceedings of International Conference on Document Analysis and Recognition, 2001, pp. 865-869.
3. Yu Zhong, Hongjiang Zhang, and Anil K. Jain, Automatic Caption Localization in Compressed Video, IEEE Transactions on Pattern Analysis and Machine Intelligence, 22, (4) (2000) 385-392.
4. Karin Sobottka, Horst Bunke and Heino Kronenberg,"Identification of Text on Colored Book and Journal Covers", Document Analysis and Recognition, 20-22 ,1999.
5. Xin Zhang, Fuchun Sun, Lei Gu, "A Combined Algorithm for Video Text Extraction",2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery, 2010.
6. X. Liu and J. Samarabandu, "An edge-based text re-gion extraction algorithm for indoor mobile robot navi-gation," inProc. of the IEEE International Conference on Mechatronics and Automation (ICMA 2005), Niagara Falls, Canada, July 2005, pp. 701–706