Comparison Of Some Text Extraction Methodologies

V. K. Yeotikar
Department Of Computer Science, Ssesa's Science College, Congress Nagar, Nagpur.

M. P. Dhore
Department of Computer Science, SSESA's Science College, Congress Nagar, Nagpur.

Abstract

In Document Image analysis the digitized images of printed documents typically consist of a mixture of text, graphics, and image elements. For proper processing and efficient representation, these elements have to be separated. For most of the applications it is essential to separate between text and non-text, because text captures the most information. These text lines may have different orientations or the text lines may be of curved shapes. Some of the techniques proposed for text string extraction are completely independent from text orientation and may deal with text in various font styles and sizes. There are many fast and efficient methods for extracting graphics and text paragraphs from printed document. This paper outlines the comparisons of some text extraction techniques proposed by researchers.

1. INTRODUCTION

The aim of document image analysis is to transform the information of a digitized document image into an equivalent symbolic representation. In most applications, text parts are the main information carrier. For that purpose, it is necessary to locate text objects within the image, recognize them, and finally extract the hidden information. Because of the increasing number of comfortable publishing systems, documents nowadays contain in addition to text also graphics and images which overlap each other. Additionally, text lines are not only horizontally aligned. Therefore finding text parts is not a trivial task.[1]

The document classification determines its document type based on its structure and then extracts the most important content based on its logical structure. A common way is to create a new form to represent the information contained in the original document based on the document structures. The usefulness and efficiency of a text processing system can be improved greatly by converting normal text representations into a new form adapted for better computer manipulation. However, most document classification systems in practice are restricted to their specific application domain. To generate a system that can be used in many different domains, the maximum use of document structures i.e. layout structure and logical structure in analyzing documents and the embodiment of the learning ability in the system allow the system to be adapted easier into a new domain.[2]

The retrieval of text information from printed color documents, such as book or journal cover pages, has gained increasing importance in recent years. To perform automatic indexing and classification of textual information, text elements must be extracted from the documents first. Due to the characteristics of an optical scanner, scanned documents contain far more colors than the original printed document.[3]

In optical character recognition (OCR), the text lines in a document must be located before recognition. This task involves correction of document skew, separation of text and graphics, and extraction of text lines. When the text lines in a document image are parallel to one another (single oriented document), simple techniques like projection profile, Hough transform, component nearest neighbor clustering method etc., are good enough to extract them. There are many artistic documents, such as advertisement, poster, or graphic illustration where text lines are not single oriented. These text lines may be multi-oriented or may be curved in shape.[4]

A considerable portion of the text on the World Wide Web is embedded in images. Moreover, a significant fraction of this image text does not appear in coded form anywhere on the page. Such HTML documents are currently not being indexed properly - image text is not accessible via existing search engines. To make this information available, techniques must be developed for assessing the textual content of Web images. One of the difficulties to be faced in detecting and extracting text embedded in images is the prodigious use of color combined with complex backgrounds. Text in WWW images can be of any color and is often intermingled with differently-colored objects in the background.[5]

Characters in the natural scenes contain a lot of information which is useful for people's daily life, such as the names of streets and shops, road signs, traffic information, etc. For travelers who are not familiar of their locus, this information is extremely important. Recently, some researchers have proposed a concept of "Smart Camera" or "Information Capturing Camera". The Smart Camera is a machine that can automatically perform text region location, character extraction and recognition in the natural scene images captured, then understand and translate the recognition results to the language desired.[6]

Portable cameras are Ubiquitous. Either in standalone versions, or incorporated in cell phones, the quality of the images has risen at a fast pace while their price has dropped drastically. Such pervasiveness has given rise to unforeseen application such as using portable cameras for digitalizing documents by user of many different professional areas for instance, students are taking photos instead of taking notes. This new research is evolving fast in many dimensions. Recent studies in the field of computer vision and pattern recognition show a great amount of interest in content retrieval from images and videos. With the help of digital camera we can capture characters and documents anywhere in the 3D environment like signs and bill-boards, color, texture, shape, as well as the relationship between them. CBDA is required because we are no longer constrained to traditional 2D images. As stated by Jung, Kim and Jain, text data is particularly interested because text can be used to easily and clearly describe the contents of an image since text data can be embedded in an image or video in different font styles, sizes, orientations, colors and against a complex background.

2. TEXT EXTRACTION

In most applications, text parts are the main information carrier. For that purpose, it is necessary to locate text objects within the image, recognize them, and finally extract the hidden information. Because of the increasing number of comfortable publishing systems, documents nowadays contain in addition to text also graphics and images which overlap each other. Additionally, text lines are not only horizontally aligned. Text isolation and extraction from varied backgrounds is a difficult problem where regional properties of text and non-text regions are used to separate them. Font, style, size and color of characters play an important role in extracting the information from document image. Digitized images of printed documents typically consist of a mixture of text, graphics, and image elements. For proper processing and efficient representation, these elements have to be separated. Working with a wide variety of documents such as newspapers, magazines, and journals, one has to extract all possible information which indicate the type of information contained in the document. Locating text image blocks and tables, and defining appropriate algorithm for text extraction is the major challenge.

3. TEXT EXTRACTION TECHNIQUES

3.1 Literature Review

A number of techniques for processing mixed mode documents are proposed in the literature.

In [Wahl,Wong,Casey] and [Wang & Srihari] approaches are presented that classify areas of a document as text, graphic, or image. The basic objects are blocks which do not overlap each other and contain only one mode of information. Some elementary image processing techniques are described in [Bartneck] classifying single image points. The aim of this approach is to reduce image noise and therefore separate noise from text.

A rule-based approach is introduced in [Fisher, Hinds]. Connected components of a smeared image are classified and grouped to words, lines, and blocks. Fletcher and Kasturi proposed another technique where eight-connected block components are the primitives. After filtering very small and large objects, a Hough transformation is performed to non-filtered objects, grouping objects associated with a particular line. All objects belonging to a line are grouped to strings and designated as text.

The systems described above use different input and produce different results. They operate quite well for specific document classes. But there is no system dealing with low quality printings where characters are split in several connected components and considering normal as well as inverse text in one analysis phase.

3.2 Mixed Mode Technique

This technique presents an algorithm for text string extraction within mixed-mode documents being able to process types of document.

The algorithm introduced separates text from non-text objects where text objects are grouped to strings. It has similarities to a few techniques proposed in literature. But it differs in two important aspects from all approaches: we are able to handle inverse text and at some critical analysis steps we are fault-tolerant. Therefore, not every preceding analysis step has to produce optimal (intermediate) results. Instead, the combination of the phases guarantees the quality of the approach, because mistakes of pre-ordered phases may be corrected.

The analysis is divided into six phases:

- Connected Component Analysis
- Filtering
- Neighborhood Determination
- String Generation
- Inverse Filtering
- Assessment.

3.3 Directed Weight Graph Technique

This method first introduces a new segmentation methodology, then describes the document's layout structure representation in terms of a directed weight graph (DWG). A three-level matching methodology is presented for document classification. After identifying the document into type, the given document finds a mask from the document sample base, and therefore the extraction can be done automatically. Finally, it introduces a way to train the system by enhanced learning algorithm.

The task of segmentation is to separate the original document image file into several rectangular areas, also called blocks. A block is the smallest unit of a maximal homogenous area, such as text, graphics, image, etc. In reality, the line spacing is an important criterion, which indicates the closeness between two adjacent blocks. However, the line spacing cannot be used as the only criterion for the segmentation because users may adjust frequently the spacing of the document based on his/her own aesthetic opinion or particular purpose, such as emphasizing a part of the content of this document. Therefore, a general segmentation only based on the physical information is not good enough to be used later by the other components of the document processing. It is essential to analyze their logical closeness. From the output of the Optical Character Recognition (OCR), we can obtain the rich text and separated blocks. The rich text information including the text with its font and attributes can be used to conduct further analysis on the text and to find the "structured" parts of the document.

The automatic segmentation component involves grouping of all the lines with the same font and the even line intervals as much as possible to form the initial segmentation. For the adjacent areas, the logical association analysis involves Combining the lines to form a region if they are of the same format; and Combining the adjacent free-text regions into one. For the IMAGE_BLOCK and GRAPHICS_BLOCK, besides compressing for the image and graphics themselves, the caption is captured and a user-defined description is also stored in order to make the later retrieval faster.

This method introduces an automated document analysis and understanding system.

3.4 Histogram-Based Clustering Technique

Many methods for color clustering have been described in the literature. For example, the histogram-based clustering method proposed computes a histogram of all colors of the input image and builds clusters by comparing the entries of adjacent cells. The fundamental idea of histogram-based clustering and its application to grey level images is briefly reviewed.

Histogram-based methods are widely used in image analysis. For the purpose of completeness, the basic steps in the clustering algorithm proposed in [3] can be summarized as follows:

1. First the histogram for a chosen feature in the image (e.g. intensity) is built. In order to reduce the

amount of data, the range of values of the feature is quantized into cells.

2. For each cell in the histogram a pointer to its largest neighbor, i.e. the neighboring cell with most of data Points, is created. If both neighbors are equal, but larger than the actual cell, the left neighbor is chosen. If both neighbors are smaller than the actual cell, no pointer is stored. If both neighbors are equal to the actual value, a pointer to the neighbor encountered first is installed.
3. After step 2 has been completed the histogram contains chains of cells pointing to a local maximum. The set of all cells belonging to such a chain builds a cluster. A unique label is attached to each cluster, and each cell of a cluster gets assigned the corresponding label.
4. Finally, the image is segmented using the clusters obtained in steps 1 through 3, i.e., a cluster label is assigned to each pixel of the image.

3.5 Clustering Technique

This technique describes a text detection algorithm which is based on color clustering and connected component analysis. The algorithm first quantizes the color space of the input image into a number of color classes using a parameter-free clustering procedure. It then identifies text-like connected components an each color class based on their shapes. Finally, a post-processing procedure aligns text-like components into text lines. Experimentally this approach is promising despite the challenging nature of the input data.

Conceptually, text extraction algorithm follows a paradigm similar to Zhong et al.'s first method. Assumes that the foreground color is roughly uniform for a given character. In other words, the color of a single character varies slowly, and its edges are distinct. Like Zhong et al.'s method, First quantize the color space of the input image into color classes by a clustering procedure. Adopt a clustering method which is intuitive and parameter-free.

Pixels are then assigned to color classes closest to their original colors. For each color class, the shape of its connected components is analyzed and classified as character-like or non-character-like. Finally, a post-processing procedure performs a "clean-up" operation by analyzing the layout of character components.

4. COMPARATIVE ANALYSIS

In this section, different techniques discussed earlier are analyzed from comparison point of view.

The first technique is Mixed-Mode, it consist of effective algorithm for string extraction within mixed-mode documents. It uses connected components as basic elements and groups them to strings considering their relations to neighboring. The main advantage of this approach is that it is able to consider both black and white characters in one and the same analysis pass. This system can be influenced by a few parameters. These parameters were initialized with standard values determined during implementation. Tests show that there is no need for manually adapting them to specific input documents. An interesting feature of this system is that it is able to extract curved text by approximation with short text strings.

The second technique is Directed Weight Graph, which describes a domain-independent automatic document image understanding system with learning ability. A segmentation method based on the "logical closeness" is proposed. A novel and natural representation of document layout structure - directed weight graph (DWG) is described. To classify a given document, a string representation matching is applied first instead of comparing with all the sample graphs. Frame template and document type hierarchy (DTH) are used to represent document logical structure and the hierarchical relation among these frame templates respectively. It consist of two methodologies of learning – learning from experience and enhanced Preceptor learning.

The third technique is histogram-based clustering, which involves experimental evaluation of two algorithms using a database with known ground truth. Both algorithms have shown good results. In particular it has turned out that the four-dimensional clustering algorithm has the potential of improved identification rate in text extraction from complex color documents. In principle, both algorithms are not restricted to document images, but can also be used for text extraction from Web pages also.

The fourth technique is clustering. It is based on color clustering followed by a connected component analysis. The algorithm works reasonably well for given the complexity of the input data, suggesting that such techniques could prove useful in Web-based information retrieval applications. It must concentrate on improving its robustness. A related issue that deserves mention is the recognition of WWW image text. This would be a difficult problem even if the text could be located perfectly. One reason for this is that it is typically rendered at a low spatial resolution (72 dpi). Initial results suggest that this may be a promising approach as well.

5. CONCLUSION

In this paper, we presented a comparative analysis of four different text extraction techniques. The first technique is Mixed-Mode, it consist of effective algorithm for string extraction within mixed-mode documents. The second technique is Directed Weight Graph. It describes a domain-independent automatic document image understanding system with learning ability. The third technique is histogram-based clustering. In this technique two algorithms have been experimentally evaluated using a database with known ground truth. The fourth technique is clustering. It is based on color clustering followed by a connected component analysis. The extraction of text from the document images has a wide range of applications and a variety of algorithms are developed for number of applications, but no algorithm guaranties cent percent performance. Extraction of text is concerned with recovering the defined attributes obscured by imperfect measurements. To represent a character class, either a prototype or a set of samples must be known. The feature selection process attempts to recover the pattern attributes characteristic of each class. Global features, such as the number of holes in the character, the number of concavities in its outer contour, and the relative protrusion of character extremities, and local features, such as the relative positions needs to be explored at large.

6. REFERENCES

[1] Frank Hones, Jiirgen Lichter. "TEXT STRING EXTRACTION WITHIN MIXED-MODE DOCUMENTS" 1993 IEEE.

[2] Xuhong Li ,Peter A. Ng."A DOCUMENT CLASSIFICATION AND EXTRACTION SYSTEM WITH LEARNING ABILITY "

[3] T. Perroud, K. Sobottka, and H. Bunke "Text extraction from color documents - clustering approaches in three and four dimensions" 2001 IEEE.

[4] Jiangying zhou, Daniel Lopresti "EXTRACTING TEXT FROM WWW IMAGES".1997 IEEE

[5] Xuewen Wang "CHARACTER EXTRACTION AND RECOGNITIONS IN NATURAL SCENE IMAGES" 2001 IEEE.

[6] F. Leabourgeois, Z. Bublinski and H. Emptoz"A Fast and Efficient Method for Extracting Text paragraphs and graphics from unconstrained Document"1992 IEEE.

[7] U. Pal and Partha Pratim Roy "Multioriented and Curved Text Lines Extraction From Indian Documents" 2004 IEEE