Study of Text Recognition from Images

Miss Tejaswini V. Ramekar & Prof. Dr. A. S. Alvi
Department of Computer Science & Engineering
Prof. Ram Meghe Institute of Technology & Research, Badnera, Amravati University

Abstract: Text recognition in images is a one of such a research area which is used to enhance a computer system with the ability to automatically read the text from images. Now-a-days there is a big demand in storing the information in paper documents format in to a computer storage disk and later reusing this information by searching process. One easy way to store the information from these paper documents in to computer system is to first scan the documents and then store them as images. It is very hard to reuse this information, by reading the individual contents and searching the contents form these documents word-by-word and line-by-line. The challenges involved in this are the font of characters in paper documents and quality of images. Due to these challenges, computer is not capable to recognize the characters while reading them. Thus there is a requirement of character recognition mechanisms to perform Document Image Analysis (DIA) which converts documents in paper format to electronic format. In this paper we have analyzed & reviewed different methods for text recognition from images.

Keywords: Document Image Analysis(DIA), electronic format, storage disk, font characteristics, text recognition

1. Introduction

As we know that we have number of newspapers and books which are in printed format related to different subjects but now-a-days, there is huge demand for the software systems to recognize characters in computer system, when information is scanned through paper documents. These days there is a growing demand in “storing the information in these paper documents in to a computer storage disk and later reusing this information by searching process”. One easy way to store information from these paper documents in to computer system is to first scan the documents. Whenever we scan the documents through the scanner, the documents are stored as images format in the computer system. These images containing text cannot be edited by the user. It is very hard to reuse this information, by reading the individual contents and searching the contents form these documents word-by-word and line-by-line. The reason for this difficulty is the font of characters in paper documents are different from font of the characters in computer system. As a result, computer is not capable to recognize the characters while reading them. This mechanism of storing the contents of paper documents in computer storage and then reading and searching the content is called document processing. Sometimes in this document processing, there is a requirement of processing the information that is related to languages other than the English in the world. This process is also called Document Image Analysis (DIA). Thus our requirement is to develop some text recognition algorithm to perform Document Image Analysis (DIA) which converts documents in paper format to electronic format.

2. Text recognition system

In this section we will study the overall architecture of Text recognition system. A Text recognition system takes image as a input which contains some text information. The output of this system is in electronic format i.e. text information in image are stored in computer readable form. The Text recognition system can be classified in following module:

A. Pre-processing Module
B. Text Recognition Module
C. Post-processing Module

A. Pre-processing Module

Generally, the paper document is scanned by the optical scanner and transform in to the form of a picture. A picture is the combinations of picture elements are also known as pixels. At this stage, we have the data in the form of image and this image can be further verified so that the important information can be regain. So to enhance the quality of input image, we perform some operation on it such as

a) Normalization

Normalization is one of the important pre-process for text recognition. The normalization is done to obtain characters of uniform size, slant and rotation.

b) Noise removal
Noise removal is one of the most significant processes. Due to this, the quality of the image will enhance and it will effect the recognition process for better text recognition in images. And it results in the generation of more accurate output at the end of text recognition processing.

**Fig.1: Architecture of text recognition**

B. Text Recognition Module

For recognising the text in output image of Pre-processing model, this module can be used and give output data which are in computer understandable format. Hence in this module following techniques are used.

a) Segmentation
In text recognition module, segmentation is done to make the separation between the individual characters of an image.

b) Feature Extraction
Feature extraction is the process of regaining the most important data from the unrefined data. The most important data is the data in which characters can be represented accurately. To store the different features of a character, the different classes are made. There are many techniques used for feature extraction like Principle Component Analysis (PCA), Independent Component Analysis (ICA), Linear Discriminate Analysis (LDA), Histogram, Chain Code (CC), Gradient Based features, etc.

c) Classification
The classification is the process of verifying each character and assigning it to the correct character class, so that texts in images are converted to computer understandable format. This process used extracted feature of text image for classification i.e. input to this stage is output of the feature extraction process. Classifiers compare the input feature with stored pattern and find out best matching class for input. There are many techniques used for classification such as Artificial Neural Network (ANN), Support Vector Matching (SVM) Template Matching etc.

C. Post-processing Module

In this module, we get the output of Text Recognition Module is in the form of text data which is understandable by computer. We can store it in some proper format (i.e. txt or MS-Word) for further use such as editing or searching in that data.

3. Literature review

We have seen some approaches for text segmentation, feature extraction and classification have been proposed in the past. In this section we discuss about that approaches which is useful for text recognition. Yang et al. [1] has proposed a novel adaptive binarization method based on wavelet filter which shows comparative performance to other similar methods and processes faster, so that it is more convenient for real-time processing. Gur et al. [2] has discussed that text recognition and retrieval is a well known problem. In this paper the authors suggest a novel text recognition algorithm based on usage of truth values relying on statistical data of the analyzed font. Automated optical character recognition (OCR) tools do not provide a complete solution and in most cases human inspection is necessary. Then new approach is defined which combines letter statistics and correlation coefficients in a set of fuzzy based rules, possibly the recognition of distorted letters that may not be regained otherwise.

Shrey Dutta, Naveen Sankaran et al. [3] has presented a novel recognition approach that results in a 15% fall in word error rate on heavily degraded Indian language document images. Naveen Sankaran and C.V Jawahar et al. [4] has propose a recognition scheme for the Indian script of Devanagari. Recognition accuracy of Devanagari script is not yet comparable to its Roman counterparts. This is due to the complexity of the script, writing style etc. Our solution uses a Recurrent Neural Network known as Bidirectional Long-Short Term Memory (BLSTM). Rhead et al. [5] has considered real world UK number plates and relates these to ANPR. It considers concepts of the significant laws and standards when applying them to real world number plates. The various manufacturing techniques and specifications of component parts are also noted. Pradeep et al. [6] have given neural network based classification of handwritten character recognition system. Each
individual character is resized to 30 X 20 pixels for processing. They are using binary features to train neural network. In post processing stage, recognized characters are converted to ASCII format. For each group new feature is calculated to discriminate characters within the group. Badawy, W. et al. [7] has discussed the Automatic license plate recognition (ALPR) is the extraction of vehicle license plate information from an image or a sequence of images. The extracted information can be used with or without a database in many applications, such as electronic payment systems and arterial monitoring systems for traffic surveillance. The ALPR uses either a black & white or color, or infrared camera to take images. Malakar et al. [8] has described that extraction of text lines from document images is one of the important steps in the process of an Optical Character m Recognition (OCR) system. In case of handwritten document images, presence of skewed, touching or overlapping text line(s) makes this process a real challenge to the researcher. The present technique extracts 87.09% and 89.35% text lines successfully from the said databases respectively.

4. Application

It is possible to apply text recognition technology throughout the whole spectrum of industries, revolutionizing the document management process. This technology permit scan documents to become more than just image files, converting into fully searchable documents with text content that is recognized by computers. Now-a-days, people no longer need to manually retype important documents when entering them into electronic databases with the help of this technology. Instead, Text recognition system extracts related information and enters it automatically. The result is approximate, efficient information processing takes a few time. In the following, we will see some applications of text recognition system,

A. Legal

Now-a-days, legal industries are working on digitizing paper documents. Documents are being scanned and entered into computer databases, in order to save space and eliminate the need to separate through boxes of paper files. Image text recognition further simplifies the process by making documents text-searchable, so that they are easier to locate and work with once in the database. Legal professionals now have fast, easy access to a huge library of documents in electronic format.

B. Banking

The uses of image text recognition is different according to the field. One widely known application is in banking, it is used to process checks without human participation. A check can be inserted into a machine, the writing on it is scanned instantly, and the correct amount of money is transferred. Now-a-days, this technology is going in perfection for printed checks, and is fairly accurate for handwritten checks as well, though it sometimes needs manual confirmation. However, this reduces wait times in many banks.

C. Healthcare

Image text recognition technology is useful in healthcare to process paperwork. Healthcare professionals always have to deal with large no. of forms for each patient, including insurance forms as well as general health forms. To keep the track of this whole information, it is useful to input related data into an electronic database that can be accessed as per our requirement. By using image recognition technology they are able to extract information from forms and put it into databases, so that every patient's data is accurately recorded. As a result, healthcare providers can offer the best possible service to every patient.

D. Image text recognition in Other Industries

Image text recognition technology is widely used in many other fields, including government agencies, education, and finance. This technology has made infinite texts available online, saving money for students and permitting knowledge to be shared. Image text recognition technology simplifies data collection and analysis, among other processes, in government agencies and independent organizations, Invoice imaging applications are used in many businesses to keep track of financial records and save a backlog of payments from piling up. As there is growing development in this technology, more and more applications are found for technology, including increased use of handwriting recognition.

5. Conclusion

In this paper, we have analyzed and reviewed different ways to find text characters from scene images. We have reviewed basic architecture of text recognition from images. In which we discussed different techniques of image processing in particular sequence for text recognition from scan image. Also, we have discussed some application of text recognition system.

References


