ISSN: 2348-4098
Volume 3 Issue 2, 2015

OPTICAL CHARACTER RECOGNITION

[1]PRIYA SINGH, [2]VIPIN GUPTA

[1]M. Tech (VLSI) Final Year, Suresh Gyan Vihar University

[2]Assistant Professor, Department of Electronics & Communication Engineering, Suresh Gyan Vihar University

ABSTRACT

The paper elaborates in detail the advantages of Optical Character Recognition based processor with that of a general purpose image processor. Application intensive processing enabled comparatively better image resolution, higher speed, and successful execution of sophisticated algorithms and so on.

However, it is a well – known fact that every coin has two sides. The same stands true for this experiment of ours as well. Apart from these numerous benefits a major problem surfaced. Because of the alignment and resolution relaed issues of the recognition system tremendous amount of noise accompanied the captured images. Careful observation concluded that Gaussian noise formed a major chunk of this bug. In this we decided that the template matching technique is the best for Gaussian noise removal. Thus we designed almost all possible types of algorithms to eliminate Gaussian noise. These algorithmss were designed on MATLAB. The codes were simulated successfully. This keeps a window of future work on this topic a definite possibility.

Thus this paper turned out to be a perfect mix of thorough study and then its subsequent implementation. This paper makes the concepts of Opical Character Recognition and its noise elimination via MATLAB crystal clear.

1. INTRODUCTION

Science fiction writers have long thought about robot with cognitive and language skills. Hollywood portrays an enchanting world of computer system and intelligence. But where do we stand today? Will computer technology continue to meet the predictions of futurists? This paper argues, in one of the many limitations of artificial intelligence; identifying characters in the document pages. This field is called optical character recognition or OCR.

This paper presents a document processing method based on the approach of optical character recognition (OCR). The text separation method of the file is maintained between them by reconstructing the characters from the original text of the page image. Seperation of text and its extraction works on the phenomenon of hierarachical framing. The process begins with framing of a single character. Once the characters are recognized, the process continues with word by word framing. It ends when all the lines of text are framed.

A powerful combination of character recognition and identification system should be able to filter out the noise and maintain the accuracy by adapting to other acoustic conditions. Designing a robust character recognition algorithm is a complex task that requires detailed knowledge of signal processing and statistical modeling. This paper demonstrates the use of MATLAB to develop algorithms for character recognition system and its workflow. On the other hand, it works similar to the human visual system to make a global file, image and text recognition without going into the details. The purpose of this study is to provide an introduction to researchers in this field. It will also discuss the level of development of this approach, progress, issues and future challenges.

Document page may contain machine-printed characters (such as this page), hand-printed or handwritten cursive text. Here, we focus on recognizing handwritten and machine-printed characters in the easiest possible way.

2. OCR SYSTEM DESIGN

Machine-printed text recognition system originated in the late 1950s and has been widely used since early 1990s in desktop computers. Many of the world's information is held in hard copy documents. OCR system releases this information via text on paper through an electronic form. Once in this form, the information retrieval system can be used to locate matter of interest, and a word processing software can be used for editing the text. OCR technology has been developed so much that today's system is indeed useful in dealing with a large variety of machine-printed documents. Processing a cleanly-printed image can deliver results with an accuracy of 99% or more.

When a scanner scans a page of text into a system (i.e. a PC or a laptop), The text is saved in the form of an electronic file composed of minute dots, also known as pixels. A computer does not considers these set of pixels as text, in fact, it is considered as an image of the text.

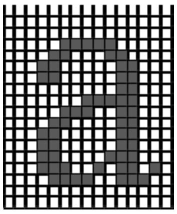

Figure 1: Image of text

These images cannot be processesd by the word processors. So, to be able to edit the group of pixels they must first be converted into words. For this, the picture

ISSN: 2348-4098
Volume 3 Issue 2, 2015

must undergo a complex phenomenon called the Optical Character Recognition. The following imae shows the general OCR system:

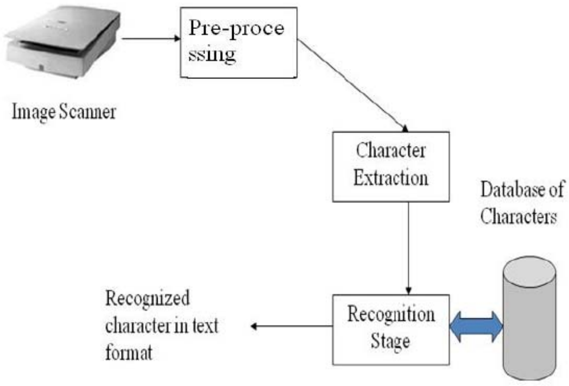

Figure 2: General OCR System

The first box in the figure corresponds to the working of the scanned image, ie, binarization, noise removal, refinement, skew correction and detection. The second box represents the second stage of preprocessing which corresponds the character extraction; it is pretreated to perform the line, word and character separation of the second stage of the project. The last phase is responsible for the feature extraction and selection resulting in image recognition.

3. WORK DONE

In this project we have performed the character recognition for both handwritten and typed characters. Lets begin with a grayscale image with handwritten characters. The first image that we considered in this project is given below:

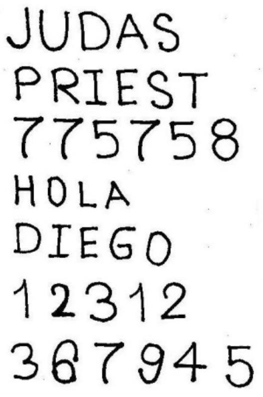
JUDAS
PRIEST
775758
HOLA
DIEGO
12312
367945

Figure 3: Input text image

This is a scanned image that consists of both alphabets and numerals. Thus, to perform the character recognition we use the template matching technique. For this, we will start with creating a template for all the alphabets from A-Z and numels from 0-9. We crop each character as compactly and closely as possible removing all the extra unwanted space as far as possible. We need to be very efficient in cropping because any unwanted area in the templates might introduce discrepancies and lead to errors in character recognition. The following figure shows the cropped numerals from 0-9.

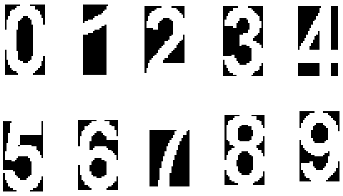

Figure 4: Cropped numerals' images for templates

Similarly, we crop all the images from A-Z. The following figure shows the cropped alphabets used as templates.

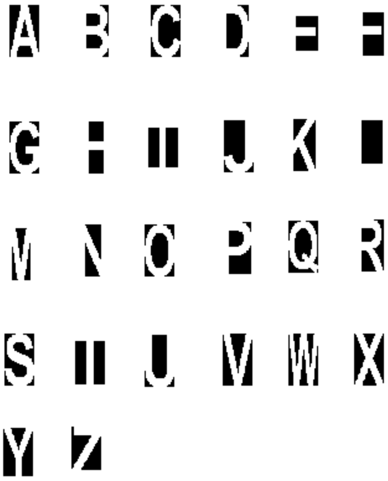

Figure 5: Cropped Alphabets used for templates

All these numbers and alphabets are saved seperately. The characters in the input image are compared to these templates and the matching algorithm is used for comparison.

Now, since the templates are created we can start with the input image processing.We start with reading the input image by using the command 'imread'. The read image is stored in 'imagen'. We then use 'imshow' command to show thw image stored in 'imagen' Then we perform the conversion of a colored (RGB) image to black and white (BW). For this we first need to convert the RGB image to grayscale image which is stored in variable 'imagen'. This grayscale image is then converted to black and white image and again stored in 'imagen'. In other words, a colored image which is also a 3Dimensional image is converted to 2 Dimensional image, this is black and white image.

After that, all those objects that are lesser than 30 pixels are removed from the input image. We have chosen 30 as the threshold value in our case study whereas this can be varied depending on the field of application. If you want the comparison process to be very stringent you can set the threshold to a higher value.

Thereafter, we load the templates created earlier to begin the template matching process. The different lines of the input image are seperated so that we can show the result in the output image line by line. Then comes the extraction process whereby the letters in the input image are extracted and stored in different variables. Process of extraction and normalization work hand in hand. This is why the normalization technique comes in place where we resize the leters extracted from the input image to match the size if the templates. The character from the scanned image is normalised from 60 X 60 pixel into 42 X 24 pixel for classification. This is a part of feature extraction which is important because both the horizontal and vertical components need to exatcly match with the templates' components for accurate results. This is also where the comparison is done to get the results.

The extracted normalised images are then converted to text by reading the letters in the images.The identified letters or alphabets are then concatenated as per the input image to giv ethe final result in a separate file. Here we have used text. Text file to show the result of character recognition. You can choose to show the output line by line or character by character or all at once. To show the result character by character we have introduced a pause of 0.5 seconds between each character so that it can be easily identified. Finally winopen command is used to open the output file to display the identified characters.

4. CONCLUSION

This study presents a template matching character recognition system, we can extract and identify the character and record the results in the scanned image. The system uses template matching alorithm and involves a feature extraction and classification techniques of normalization. In our paper, the classification and identification technology only applies to black and white images, so we color (RGB) image is converted to black and white images in the application of technology and algorithms before continuing. However, in this way, we were able to successfully identify the colored and black and white images. In our study, we set the threshold at 30 pixels, and eliminates all obects less than 30, the value can be applied which threshld area is varied.

However, errors can be introduced if the template is created in the system do not have any valid cropped. It is important to eliminate all unnecessary spaces, and to ensure that peacekeeping has always maintained is very important.

5. FUTURE SCOPE

In the future, OCR has been speculated that the use of more advanced. Some of these advances have been in operation. Most people think that is used more and more, and paper files are eliminated for reading written text will be reduced, such as electronic data interchange (EDI).

In fact, many debate whether or not the paper even EXST in the coming decades. However, recent studies have found AIIM imaging increases because of increased consumption of printing paper. While the debate, obviously, will cause OCR progress to a higher height than before.

OCR has been used to detect viruses, or unfortunately they are created, and to stop spammers. Anti-spam program continues to improve, the need to increase technology is a constant. OCR is used to prevent the virus by detecting codes hidden in the image. The difficulty of the information transfer from the old system of modern operating systems might be more trouble, OCR can read screenshots. This information can facilitate heat transfer between incompatible technologies. A current hope is to develop OCR OCR software can read the opportunity to compress the file. Be compressed into one image, and save it as ASCII text or hexadecimal data can be read using OCR and send back into readable text.

Finally, it is anticipated, OCR will be used in a more advanced development of the robot. Robot's eyes are essentially meant to enter information into a camera. If OCR is used to help the robot to comprehend the text, use may be almost endless. All of these advances, more are expected to remain in use OCR technology, continues to expand its current capabilities.

REFERENCES

1) Mohamed Cheriet, Nawwaf Kharma, Cheng-Lin Liu and Ching Y. Suen, "Optical Recognition Systems", A guide for students and practitioners, by John Wiley & sons, inc.

2) Anita V. Gawand, Prashant Lokhande, Sulekha daware, and Umesh Kulkarni, "Image Segmentation for Nature Images using K-Mean and Fuzzy C-Mean", in International Conference on Recent Trends in Information Technology and Computer Science (IRCTITCS) 2011.

3) Vijay Ranjan Nadar, "Optical Character Recognition", The Code Project Open License (CPOL), 14 Oct 2012.

4) Ruxandra Cohal, "Multiobjective Approaches in ImageSegmentationRuxandra Cohal".

5) Diego Barragán, "Optical Character Recognition", at www.mathworks.com, 09 Feb 2009.

6) Mohit Agarwal and Gaurav Dubey, "Application of clustering technique for Image Segmentation", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 4, April 2013.

7) Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine S¨usstrunk, "SLIC Superpixels", School of Computer and Communication Sciences (IC) Ecole Polytechnique F´edrale de Lausanne (EPFL).

8) Georgios Vamvakas, "Processing and Recognition of Handwritten Documents", Computational Intelligence Laboratory Institute of Informatics and Telecommunications National Centre for Scientific Research "Demokritos.

9) Elton Dunn, "How to Convert a Scanned Document to Word Format", eHow.com contributor.