SURVEY ON IMAGE PROCESSING TECHNIQUES

Nancy Aggarwal[1], Shilpa Sethi[2], Priyanka Rawat[3]
Research Scholar [1]
Assistant Professor[2]
Research Scholar [3]
[1]Department of Computer Engineering,
[1] J.C.Bose University of Science & Technology, YMCA, Faridabad, india

ABSTRACT- Optical character recognition (OCR) is process of classification of optical patterns contained in a digital image. The character recognition is achieved through pre-processing, segmentation and feature extraction. The paper starts with a brief background and history of OCR systems. The different techniques of OCR systems such as Optical Scanning, Location Segmentation, Pre-processing, Segmentation and Feature Extraction, Recognition the character are used. The different applications of OCR systems are highlighted next followed by the current status of the OCR systems. Finally, the future of the OCR systems is presented.

Index Term - Image Segmentation, Pre-processing, Tesseract, OpenCV, NumPy, OCR (Optical Character Recognition).

1. INTRODUCTION

Nowadays all over digitization technology is used. Text Recognition usually abbreviated to OCR, involves a computer system designed to translate images of typewritten text (usually captured by a scanner) into machine editable text or to translate pictures of characters into a standard encoding scheme representing them. OCR began as a field of research in artificial intelligence and computational vision. Text Recognition used in official task in which the large data have to type like post offices, banks, colleges etc., in real life applications where we want to collect some information from text written image. People wish to scan in a document and have the text of that document available in a .txt or .docx format. Tesseract image processing is the new approach to deal with the limitations of Image processing. The accuracy of Tesseract can be increased significantly with the right Tesseract image preprocessing tool chain.

Optical Character Recognition (OCR) technology got better and better over the past decades thanks to more elaborated algorithms, more CPU power and advanced machine learning methods. Getting to OCR accuracy levels of 99% or higher is however still rather the exception and definitely not trivial to achieve. At Docparser we learned how to improve OCR accuracy the hard way and spent weeks on fine-tuning our OCR engine.

When it comes to OCR accuracy, there are two ways of measuring how reliable OCR is: Accuracy on a character level In most cases, the accuracy in OCR technology is judged upon character level. How accurate an OCR software is on a character level depends on how often a character is recognized correctly versus how often a character is recognized incorrectly. An accuracy of 99% means that 1 out of 100 characters is uncertain. While an accuracy of 99.9% means that 1 out of 1000 characters is uncertain. Measuring OCR accuracy is done by taking the output of an OCR run for an image and comparing it to the original version of the same text. You can then either count how many characters were detected correctly (character level accuracy), or count how many words were recognized correctly (word level accuracy).

To improve word level accuracy, most OCR engines make use of additional knowledge regarding the language used in a text. If the language of the text is known (e.g. English), the recognized words can be compared to a dictionary of all existing words (e.g. all words of in the English language corpus). Words containing uncertain characters can then be “fixed” by finding the word inside the dictionary with the highest similarity.

The handwritten characters or typewritten from the image is convert into the machine editable format with the help of OCR (Optical Character Recognition). OCR is used to extract the text from the scanned document. OCR reads a wrong and cheap code ad estimate the best result for what the code is. OCR does not work on standard of characters. Standards includes the quality and sharpness of character. OCR include various phases like Classifications and Recognition, Feature Extraction, Segmentation, Pre-processing [1]. Phases of OCR work sequentially which shown in figure 1:-

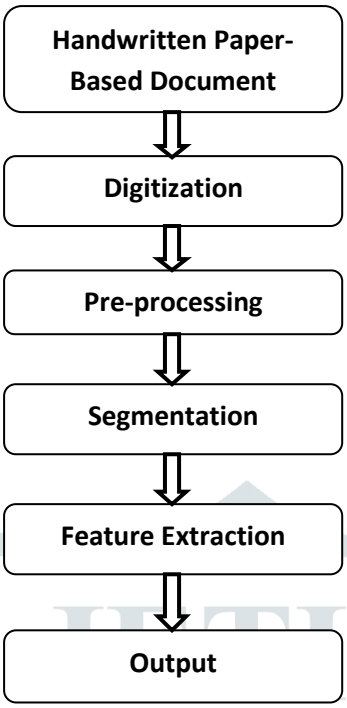

Figure 1 Work Flow of OCR

1.1 Digitization

Digitization is the process to converting handwritten paper-based document into digital format. Here, each document is organized into bits. Digitization means the analog format into a numerical format. Digitization is used various methods to scan the document and produce digital format representation of scanned document. Digitization used various scanner to scan the document and the digital representation of scanned image was going for the preprocessing phase.

1.2 Pre-Processing

The preprocessing phase, the number of operation implement on scanned image. it enhance the some features of image that are important for segmentation phase. The window sized image is extract from normalized the gray scale image. the bitmap image is produce when noise is reduce from the image then the bitmap map image is converting into the thinned image.

1.3 Segmentation

Segmentation of character is the most important phase of OCR. Segmentation is separation of individual word of an image. Segmentation of character from the printed document is difficult because of its standard format. Sometimes parts of two contiguous character are touched or over extended and the difficulty is create by these type of critical situations.

1.4 Feature Extraction

It is the process of collecting higher-level information of an image such as shape, texture, color, and contrast. In fact, texture analysis is an important parameter of human visual perception and machine learning system. It is used effectively to improve the accuracy of diagnosis system by selecting prominent features. They introduced one of the most widely used image analysis applications of Gray Level Co-occurrence Matrix (GLCM) and texture feature. This technique follows two steps for feature extraction from the medical images. In the first step, the GLCM is computed, and in the other step, the texture features based on the GLCM are calculated. Due to the intricate structure of diversified tissues such as WM, GM, and CSF in the brain MR images, extraction of relevant features is an essential task. Textural findings and analysis could improve the diagnosis, different stages of the tumor (tumor staging), and therapy response assessment.

2. METHODOLOGY

The paper proposed system for font style recognition and classification system mainly involves three main stages. In the stage 1, an input image is acquired for processing and proceeded for pre-processing and further directed for feature computation in stage two. Finally, the features computed are classified through a SVM classifier [2]. In the paper proposed a system that describe the overall architecture of Text recognition system. A Text recognition system receives an input in the form of image 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 [3].

The paper [4] proposed the use of content base image retrieval (CBIR) techniques for indexing and retrieval of handwritten documents in Thai language. Issues associated with Thai handwritten documents are the lack of spacing between words, multi-level alphabets and different writing styles. This causes low recognition rate based on automated techniques such as Optical Character Recognition (OCR). This paper also examined off-line signature recognition techniques in order to adapt to Thai handwriting system for matching data. The objective of the proposal is to develop a semiautomated method to index and retrieve Thai handwritten documents based on sampled keywords by combining CBIR and signature recognition techniques.

In this paper [12], a computer vision and character recognition algorithm for a license plate recognition (LPR) is presented to be used as a core for intelligent infrastructure like electronic payment systems (toll payment, parking fee payment), freeway. Based on the connected component analysis and novel adaptive image segmentation technique is presented [12].

The paper [8] designed a system which consider the process dedicated to the quality assessment of word recognition. This process has to be performed without comparison with ground truthed data. OCR performance estimated using support vector regression. J. r´ı Matas [10] presented an end-to-end real-time scene text localization and recognition method. In the first stage of the classification, the probability of each ER being a character is estimated using novel features calculated with O(1) complexity. In second stage only ERs with locally maximal probability are selected.

The segmentation for separation the text from the document image is makes easy with the help of projection profile-based methods. OCR used different type of method between each stage. Segmentation of text is done by the projection profile-based methods. They designed an algorithm, which used for correct the skew angles of text document [5]. In the paper [6] Many feature like translation and rotation cannot be retrieving by CBIR. So Scale Invariant Features transform or widely known as SIFT is one of the techniques that have been successfully used for local interest point detector and its descriptors. Scale-invariant feature transform (or SIFT) is an algorithm in computer vision to detect and describe local features in images.

In this paper we proposed algorithm for solving the problem of offline character recognition. We had given the input in the form of images. The algorithm was trained on the training data that was initially present in the database. We have done preprocessing and segmentation and detect the line [7]. The paper we proposed algorithm for solving the problem of offline character recognition. We had given the input in the form of images. The algorithm was trained on the training data that was initially present in the database. We have done preprocessing and segmentation and detect the line [9].

Huei-Yung Lin and Chin-Yu Hsu [11] presented neural network based approach which reduces the training time and maintains the high recognition rate. Multi-stage approach and Preprocessing Feature to be extracted Search Image Library A Matched letter Match A Image see International Journal of Computer Applications (0975 – 8887) Volume 160 – No 6, February 2017 22 pre-processing are done for the experiment. Preprocessing is performed to partition the training data prior to training stage.The review summary is shown in table 1:-

Table 1 Review Summary

Serial no.	Research Paper	Problem Statement	Solution	Advantage	Disadvantage
1.	A Font style classification system for English OCR.	Limited Number of Font Style	Provide more than 10 font style as a input using Tesseract.	It support all the font style and give the output.	Quality of the output depends on quality of the image.

2.	Using Content based Image Retrieval Techniques for the Indexing and Retrieval of Thai Handwritten Documents.	Image to Word Format	The data available on papers in to computer process able documents file (.doc, .txt) using OCR	1. The documents can be editable and reusable. 2. The processing of OCR information is fast. Large quantities of text can be input quickly. 3. A paper based form can be turned into an electronic form which is easy to store or send by mail.	1. OCR systems are expensive. 2. Images produced by scanner consume lot of memory space. 3. Images lose some quality during scanning and digitizing process.
3.	Text Recognition from image	Multiple Number of Image	We cover different techniques to improve OCR accuracy with multiple files using some logical OCR operations.	1. If one file in the computer gets error then other file not get affected and that logical OCR operation for multiple file handle this efficiently. 2. ANPR also use this system and medical diagnosis report is also use these OCR logical operations.	1. Time consuming process. 2. find the error occurrence in some file is difficult.
4.	A Different Image Content-based Retrievals using OCR Techniques.	Character Recognition of Single File with Low Accuracy	Advance Image Processing Provide The High Accuracy Of OCR Technology using Tesseract, OpenCV, NumPy, Libraries.	1. It is cheaper than paying someone amount to manually enter large amount of text data. Moreover it takes less time to convert in the electronic form. 2. The latest software can re-create tables as well as original layout.	1. All the documents need to be checked over carefully and then manually corrected. 2.OCR systems are expensive. 3. Images produced by scanner consume lot of memory space. 4. Images lose some quality during scanning and digitizing process.

3. CONCLUSION

In this paper we proposed algorithm for solving the problem of offline character recognition. We had given the input in the form of images. We have done preprocessing and segmentation and detect the line. This paper elaborated survey of disparate techniques for OCR" has been studied. Handwritten character, natural scene images, business cards and TV set images are selected for experimentation. A systematic flow of OCR system is discussed. In this paper projection profile based method for segmentation, Fourier transform technique is for preprocessing, and nearest neighbor classifier for classification is described. This paper can be helpful to the researcher for selecting most appropriate techniques to achieve optimum results for application according to a different parameter.

REFRENCES

[1] E. N. Bhatia, "Optical Character Recognition Techniques : A Review," IJARCSSE, vol. 4, no. 5, pp. 1219–1223, 2014.
[2] Bharath V, N.Shoba Rani "A Font style classification system for English OCR." International Conference on Intelligent Computing and Control (I2C2) 2017.
[3] Pratik Madhukar Manwatkar, Mr. Shashank H. Yadav "Text Recognition from Images." IEEE Sponsored 2nd International

Conference on Innovations in Information,Embedded and Communication systems (ICIIECS) 978-1-4799-6818-3/15/$31.00 © 2015 IEEE.

[4] Seksan Sangsawad, Chun Che Fung "Using Content Based Image Retrieval Techniques for the Indexing and Retrieval of Thai Handwritten Documents." IEEE Sponsored 2nd International Conference on Innovations in Information,Embedded and Communication systems (ICIIECS) 978-1-4244-5586-7/10/$26.00 2010 IEEE.

[5] A. S. Sawant, "Script Independent Text Pre-processing and Segmentation for OCR," Int. Conf. Electr. Electron. Signals, Commun. Optim. - 2015, pp. 1–5, 2015.

[6] Poonam A. Wankhede, Dr. Sudhir W. Mohod "A Different Image Content-based Retrievals using OCR Techniques." in International Conference on Electronics, Communication and Aerospace Technology ICECA 2017 978-1-5090-5686-6/17/$31.00 ©2017 IEEE.

[7] Ahmed BEN SALAH, Jean philippe MOREUX "OCR performance prediction using cross-OCR alignment." in OCR performance prediction using cross-OCR alignment 978-1-4799-1805-8/15/$31.00 ©2015 IEEE.

[8] Ali Farhat, Ali AI-Zawqari, Xiaojun Zhai "OCR Based Feature Extraction and Template Matching Algorithms for Qatari Number Plate." in 978-1-4673-8743-9/16/$31.00 ©2016 IEEE.

[9] Chowdhury Md Mizan, Tridib Chakraborty and Suparna Karmaka "Text Recognition using Image Processing." International Journal of Advanced Research in Computer Science ISSN No. 0976-5697 Volume 8, No. 5, May – June 2017.

[10] J. r´ı Matas, "Real-Time Scene Text Localization and Recognition," IEEE, pp. 3538–3545, 2012.

[11] H. Lin and C. Hsu, "Optical Character Recognition with Fast Training Neural Network," IEEE, pp. 1458–1461, 2016.

[12] C. N. E. Anagnostopoulos, I. E. Anagnostopoulos, V. Loumos, and E. Kayafas, "A License Plate-Recognition Algorithm for Intelligent Transportation System Applications,"IEEE, vol. 7, no. 3, pp. 377–392, 2006.

[13] S Sethi ad A Dixit, "An Automated User Interest Mining Techniques for Retrieving Quality Data." Journal of Business Analytics (IJBAN) 4 (2), 62-79, 2017.

[14] S Sethi ad A Dixit, "A novel page ranking mechanism based on user browsing Patterns." Software Engineering, 37-49, 2019.

[15] S Sethi ad A Dixit, "Design of personalized search system based on user interest and query structuring," 2nd International Conference on Computing for Sustainable Global, 2015.

[16] Nancy Aggarwal and S Sethi, "Automated Number Plate Recognition Using Template Matching," International Journal of Computer Science and Engineering, Volume-6, Issue-12, Dec 2018.