IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

OPTICAL AND INTELLIGENT CHARACTER RECOGNITION SYSTEM

Dr.R.Satheeskumar, Professor

Department of Computer Science and Engineering,

Narasaraopeta Institute of Technology, Andhra Pradesh ,India.

Abstract

This paper, OCR is to develop OCR software for online/offline handwriting recognition. OCR is an Optical Character acknowledgment and is the mechanical or electronic interpretation of pictures of manually written or typewritten content (more often than not caught by a scanner) into machine-editable content.OCR is a field of research in example acknowledgment, computerized reasoning and machine vision. Handwritten recognition is used most often to describe the ability of a computer to translate human writing into text. This may take in one of the two ways, either by scanning of written text or by writing directly on peripheral input devices. Now they are going to implement the software which will recognize the characters from online or offline document (in image format) and use it as individual user profile. Here they are creating OCR which will perceive manually written English characters. This system can be used by multiple users. They can do this by improving our software for recognizing the handwriting of more than one user. Likewise in the event that they can take the stroke data and offer it to our framework, at that point it will be conceivable to perceive even cursive content moreover. The recognized characters are stored in the text file. They can add words to the sound files and invoke them through the program, so that the recognized words can be read aloud. Therefore they can make the PC read the manually written report.

Key Words - OCR, Handwritten recognition, Text capture, Text Pattern recognition.

1. Introduction

The paper is about Optical Character Recognition. It is a process of classifying optical patterns with respect to alphanumeric or other characters. Optical character recognition process includes segmentation, feature extraction and classification. Text capture converts Analog text based resources to digital text resources. And then these converted resources can be used in several ways like searchable text in indexes so as to identify documents or images. As the first stage of text capture a scanned image of a page is taken. And this scanned copy will form basis for all other stages. The very next stage involves implementation of technology Optical Character Recognition for converting text content into machine understandable or readable format.

OCR analysis takes the input as digital image which is printed or hand written and converts it to machine readable digital text format. Then OCR processes the digital image into small components for analysis of finding text or word or character blocks. And again the character blocks are further broken into components and are compared with dictionary of characters. DotNet is an environment where problems and solutions can be denoted in terms of mathematical notations. A use of DotNet includes analysis, algorithm development, computation and much more.

DotNet is a system where elements are placed in an array but are not required any dimensionless. It helps us to solve our problem in no time and provides an easy solution. The OCR text is written into a pure text file that is then imported again to a search engine. The text is used as index searching of the information. Accuracy rates are measured in several ways and the ways they are measured impact the accuracy rate.

This technology is employed for a variety of applications, such as data entry of documents, automatic number plate recognition, digitization of printed documents in Google Books, and even beating CAPTCHA anti-bot systems. There are two different techniques or algorithms in optical character recognition: pattern recognition and feature extraction, and each technique are worth looking at in a little bit more detail.

1.1 Pattern recognition:

Using this technique, the computer tries to recognize the entire character and matches it to the matrix of characters stored in the software. As a result, this technique is also known as pattern matching or matrix matching. The drawback of this technique is that it relies on the input characters and the stored characters being of the same font and same scale Scan2CAD applies Neural Networks to the task of pattern matching. Neural networks work in an analogous way to the human brain. They learn to recognize shapes and patterns from a range of examples. Scan2CAD includes a feature allowing the user to train their own Neural Networks to recognize font styles unique to their drawings.

1.2 Feature extraction:

This one is a much more sophisticated way of spotting characters. It decomposes characters into “features” like lines, closed loops, line directions and intersections. Let’s take letter A as an example. If the computer sees two angled lines that meet at the top, and both lines are joined together by a horizontal line in the middle, that’s a letter A. By using rules like these, the program can identify most capital ‘A’s, regardless of the font that it is written in.

1.3 Principles of OCR Technology:

In principle, systems utilizing optical character recognition (OCR) can only recognize machine print. By making use of pattern-matching technology, OCR successfully translates patterns as well as shapes of characters generated by machines into corresponding codes for the computer. Although advanced systems can differentiate between multiple fonts, they are only able to process standard fonts - for example, Arial

and Times New Roman. When all the characters in a word are distinguished, the word is then compared to a vocabulary of probable responses for an end result.

By utilizing CVISION's OCR technology, Maestro Recognition Server, you can save time and increase your level of organization by not resorting to time-consuming techniques to find PDF files. In addition, you also have the choice of editing PDF files following the use of OCR. Hence, this also makes it easier for modifications to be made in the future. Although the OCR technology process itself can sound tedious, there are hosts of software that can provide a high level of efficiency. Some OCR software offers ongoing guidance so that users are able to understand concepts about the process quickly without much training.

2. Related Works

Ayatullah Faruk Mollah et dl. , This paper presents a complete Optical Character Recognition (OCR) system for camera captured image/graphics embedded textual documents for handheld devices. At first, text regions are extracted and skew corrected. Then, these regions are binaries and segmented into lines and characters. Characters are passed into the recognition module. Experimenting with a set of 100 business card images, captured by cell phone camera, we have achieved a maximum recognition accuracy of 92.74%. Compared to Tesseract, an open source desktop-based powerful OCR engine, present recognition accuracy is worth contributing. Moreover, the developed technique is computationally efficient and consumes low memory so as to be applicable on handheld devices.

Abdul Rahiman M et al. India is a multilingual and multi-script country where a line of a bilingual document page may contain tex words both in regional language and in English. Recognition of documents containing multi-scripts is really a challenging task, which needs more effort of the OCR designers for improving the accuracy rate. This paper presents a Bilingual OCR system for printed Malayalam and English text. Here we propose an algorithm which can accept scanned image of printed. Characters as input and produce editable Malayalam and English characters in a predefined format as output. The image acquired is segmented into line and character-wise using pixel by pixel approach by scanning from top-left of the image to bottom-right. The character image obtained after segmentation is resized to 16 x 16 bitmap which is used for comparison. The database contains characters in various fonts of both the languages. This database is used for comparison with the resized character image. The comparison is done using pixel-match algorithm. The matched character is displayed in the notepad. An efficiency of 87.25% is obtained using this approach.

Naganjaneyulu et al., The performance of optical character recognition (OCR) algorithm is poor on low resolution scanned text images. The conventional low pass filters inL2 space can slightly improve the performance. The method of enhancement of poor resolution text images using a low pass signal filtering algorithm in the weighted Sobolev space results in high pass correction similar to un sharp masking. This can further improve the performance of OCR on low resolution text images. In this paper, the performance of a typical OCR system on low resolution scanned text images, is studied without using any preprocessing

step, with low pass filtering in L2 space, and compared with low pass filtering in weighted Sobolev space as pre processing steps.

Mehmet Yasin AKPINAR et dl., The conversion of image-based documents into digital and processible forms can be accomplished quite successfully with optical character recognition (OCR) tools. However, there are still problems with preserving the format on the original document. An important one of these problems is the reading of the tabular data. In this paper, a method is proposed in which the tabular data contents of hard-copy documents is extracted from the text and character positions which are obtained from an OCR tool and transferred to digital forms. The performance of the method is measured by the number of detected rows and columns and presented with the results of other commercial products.

Y. Tang et dl., Historical Chinese character recognition has been suffering from the problem of lacking sufficient labeled training samples. A transfer learning method based on Convolutional Neural Network (CNN) for historical Chinese character recognition is proposed in this paper. A CNN model L is trained by printed Chinese character samples in the source domain. The network structure and weights of model L are used to initialize another CNN model T, which is regarded as the feature extractor and classifier in the target domain. The model T is then fine-tuned by a few labeled historical or handwritten Chinese character samples, and used for final evaluation in the target domain. Several experiments regarding essential factors of the CNN based transfer learning method are conducted, showing that the proposed method is effective.

S.Thiyagarajan et al., Blind people are unable to perform visual tasks. The majority of published printed works does not include Braille or audio versions, and digital versions are still a minority. In this paper, the technology of optical character recognition (OCR) enables the recognition of texts from image data. The system is constituted by the raspberry pi, HD camera and Bluetooth headset. This technology has been widely used in scanned or photographed documents, converting them into electronic copies. The technology of speech synthesis (TTS) enables a text in digital format to be synthesized into human voice and played through an audio system. The objective of the TTS is the automatic conversion of sentences, without restrictions, into spoken discourse in a natural language, resembling the spoken form of the same text, by a native speaker of the language.

S. Anbukkarasi et dl, This paper The binary conversion technique is used to preprocess the scanned documents by removing the noise and with the help of horizontal profile technique the lines were segmented. Each character in the scanned page is extracted by the segmented lines. This character segmentation technique is handled by the ideology called vertical projection system. The features can be extracted and characters are classified from the extracted characters. For the classification and feature extraction Zoning and Neural network have been used respectively. Finally the recognized characters are converted into editable text with an accuracy of 87%.

3. Proposed Solution

Content based image retrieval of images from a database that are similar in visual content to a query image. The basis of CBIR is to extract and index some visual features of the images such as Color, Shape, Texture to search user required from large image database according to user requests in the form of a query. The proposed system deals the results to demonstrate the pixel size in 384×256 pixels. The presented system algorithm is much efficient image searching techniques for nearest value from the database and that provides a new feature extraction. Established on images recapture which calculate the match of each image in its data gather. To develop and put into a practice and efficient feature extraction Modified K-Nearest Neighbor and Support Vector Machine. To extract feature according to the retrieval content of the images from the datasets. This approaches tested on Corel database (corel-100k). Besides, the image retrieval system achieving a significant speed up in query time while providing competitive retrieval accuracy.

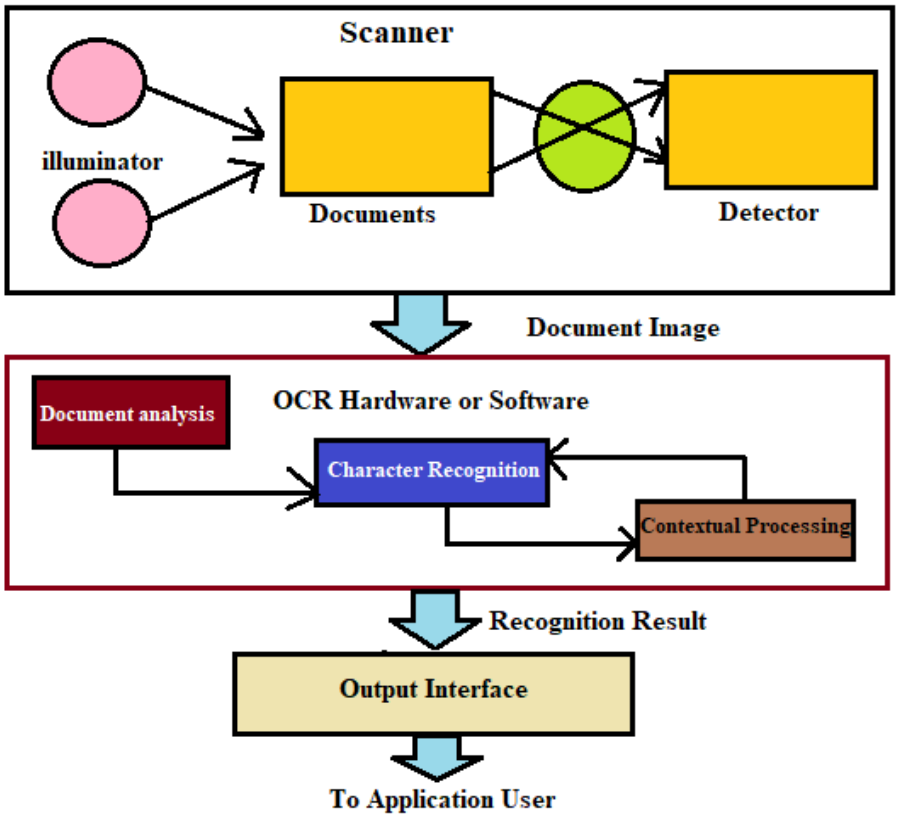


Fig 3.1: Architecture Diagram of Proposed

3.1 Document Processing:

Data Processing is accessed by administrator whose role is our application is a system admin. This module perform certain activities such as scanning documents, storing them as images, recognizing characters in images to transfer them into word format. The module supports the following services:-

- Scanning printed documents
- Storing the documents as snapshots or images
- Processing those image-based documents
- Converting these image-based documents into e-documents
- Recognizing the characters in documents

3.2 System Training:

System Training can be accessed by both the administrator and the end user. Before converting the printed documents in to editable and searchable documents, the first and the mandatory step is providing training to the system. Here training in the sense the font followed in the scanned document should be identified by the user. Then the user types all the characters that required for recognition from the scanned document as an image file. This image file should be provided as an input during the training process. This module supports,

- Training the system with pre-defined fonts.
- Training the system with the new fonts that are not present in the system and that cannot be identified by the system.

3.3 Document recognition:

Document recognition can be accessed by both the administrator and the end-user. Once the printed documents are converted into structured documents, any user can recognize the characters present in the document. That means the user canrecognize the characters of any language he chooses which makes OCR moreflexible. This is the module where the main functionality of OCR is tested. Under this module, there are two types of recognition. They are handwritten recognition and scanned document recognition.

In handwritten recognition, the handwriting of the user in any language is trained to the system only for the first time. From there on-wards, the system recognizes the characters or words written by the user. Thus handwritten document recognition In scanned document recognition, the system is first trained with the font characters in the document in the training module itself. Now in the recognition module, the system takes the scanned documents image as an input file, first crops the image and then extracts/recognizes the characters from the document and makes these documents editable and searchable. Thus the scanned document recognition recognizes the characters from the scanned document image and makes the document editable and searchable. Hence the document recognition module on a whole supports the following services,

- Converts the documents into specific format
- Recognizes the characters
- Heterogeneous character Recognition

3.4 Document Editing:

Document Editing can be accessed by both the administrator and the end-user during document editing to implement the character recognition process. Once the scanned documents are stored, they reside in computer memory. This data resides in the form of an image that is just viewable in an image viewer. Hence the documents may be MS-word, Text, as specified by the user. The objective of this module is,

- Addition of specific content to the documents
- Deletion of certain content from documents
- Any other modification of document.

3.5 Document Searching and Save

This process can be accessed by both the administrator and the end-user during the search of the user required document to implement the character recognition process on it. The user requests the system to search for a particular document. Then the system finds the documents based on OCR methodology and returns the result of the search to the user.

3.6 Optical Language Symbols

Several languages are characterized by having their own written symbolic representations (characters). These characters are either a delegate of a specific audio glyph, accent or whole words in some cases. In terms of structure world language characters manifest various levels of organization. With respect to this structure there always is an issue of compromise between ease of construction and space conservation. Highly structured alphabets like the Latin set enable easy construction of language elements while forcing the use of additional space. Medium structure alphabets like the Ethiopic (Ge'ez) conserve space due to representation of whole audio glyphs and tones in one symbol, but dictate the necessity of having extended sets of symbols and thus a difficult level of use and learning. Some alphabets, namely the oriental alphabets, exhibit a very low amount of structuring that whole words are delegated by single symbols. Such languages are composed of several thousand symbols and are known to need a learning cycle spanning whole lifetimes.

3.7 Symbol image detection

The process of image analysis to detect character symbols by examining pixels is the core part of input set preparation in both the training and testing phase. Symbolic extents are recognized out of an input image file based on the color value of individual pixels, which for the limits of this paper is assumed to be either black RGB (255,0,0,0) or white RGB(255,255,255,255). The input images are assumed to be in bitmap form of any resolution which can be mapped to an internal bitmap object in the Microsoft Visual Studio environment. The procedure also assumes the input image is composed of only characters and any other type of bounding object like a border line is not taken into consideration.

3.7.1 Determining character lines

Enumeration of character lines in a character image ('page') is essential in delimiting the bounds within which the detection can proceed. Thus detecting the next character in an image does not necessarily involve scanning the whole image all over again.

3.7.2 Detecting Individual symbols

Detection of individual symbols involves scanning character lines for orthogonally separable images composed of black pixels.

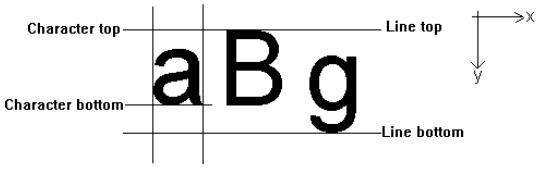


Fig 3.2. Line and Character boundary detection

From the procedure followed and the above figure it is obvious that the detected character bound might not be the actual bound for the character in question. This is an issue that arises with the height and bottom alignment irregularity that exists with printed alphabetic symbols. Thus a line top does not necessarily mean top of all characters and a line bottom might not mean bottom of all characters as well.

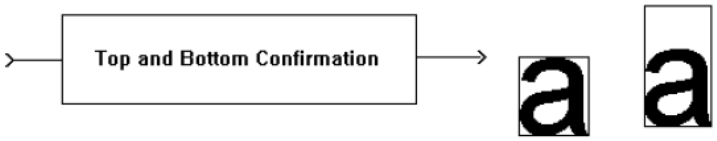


Fig 3.3 Confirmation of Character boundaries

3.7.3. Symbol Image Matrix Mapping

The next step is to map the symbol image into a corresponding two dimensional binary matrix. An important issue to consider here will be deciding the size of the matrix. If all the pixels of the symbol are mapped into the matrix, one would definitely be able to acquire all the distinguishing pixel features of the symbol and minimize overlap with other symbols. However this strategy would imply maintaining and processing a very large matrix (up to 1500 elements for a 100x150 pixel image). Hence a reasonable tradeoff is needed in order to minimize processing time which will not significantly affect the separability of the patterns. The paper employed a sampling strategy which would map the symbol image into a 10x15 binary matrix with only 150 elements. Since the height and width of individual images vary, an adaptive sampling algorithm was implemented.

Fig. 3.4 Mapping symbol images onto a binary matrix

In order to be able to feed the matrix data to the network (which is of a single dimension) the matrix must first be linearized to a single dimension. Hence the linear array is our input vector for the MLP Network. In a training phase all such symbols from the trainer set image file are mapped into their own linear array and as a whole constitute an input space. The trainer set would also contain a file of character strings that directly correspond to the input symbol images to serve as the desired output of the training.

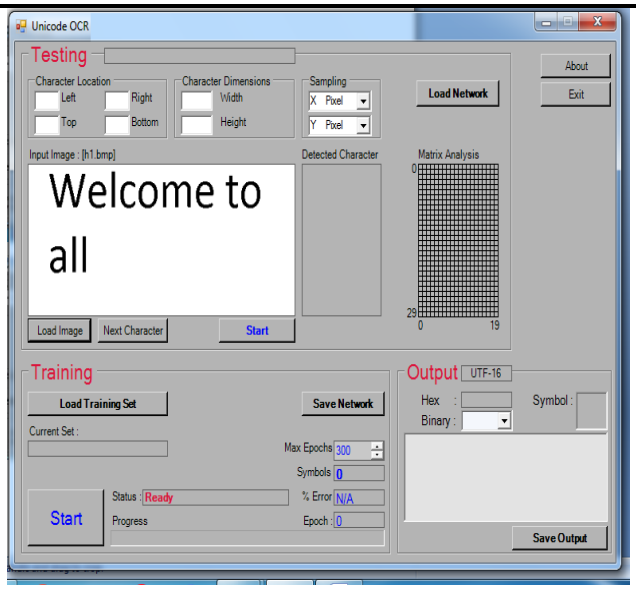
4. RESULT

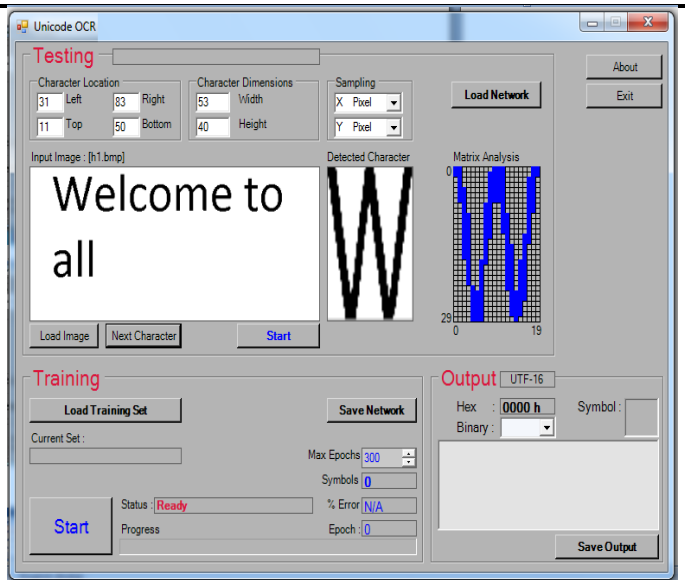
The network has been trained and tested for a number of widely used font type in the Latin alphabet. Since the implementation of the software is open and the program code is scalable, the inclusion of more number of fonts from any typed language alphabet is straight forward. The necessary steps are preparing the sequence of input symbol images in a single image file (*.bmp [bitmap] extension), typing the corresponding characters in a text file (*.cts [character trainer set] extension) and saving the two in the same folder (both must have the same file name except for their extensions). The application will provide a file opener dialog for the user to locate the *.cts text file and will load the corresponding image file by itself. Although the results listed in the subsequent tables are from a training/testing process of symbol images created with a 72pt. font size the use of any other size is also straight forward by preparing the input/desired output set as explained. The application can be operated with symbol images as small as 20pt font size.

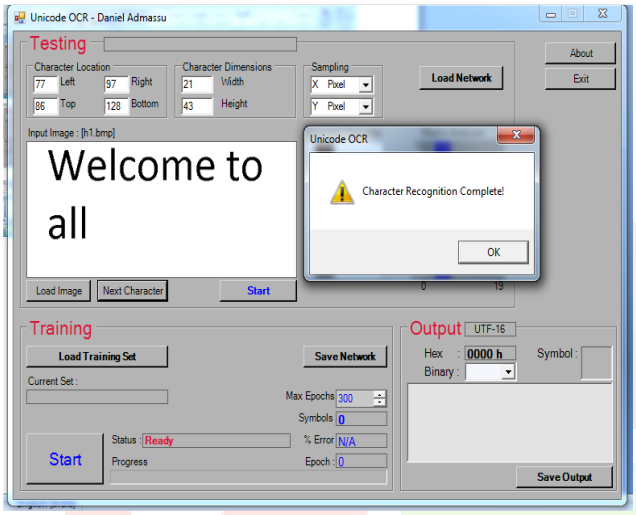
A. Results for variation in number of Characters

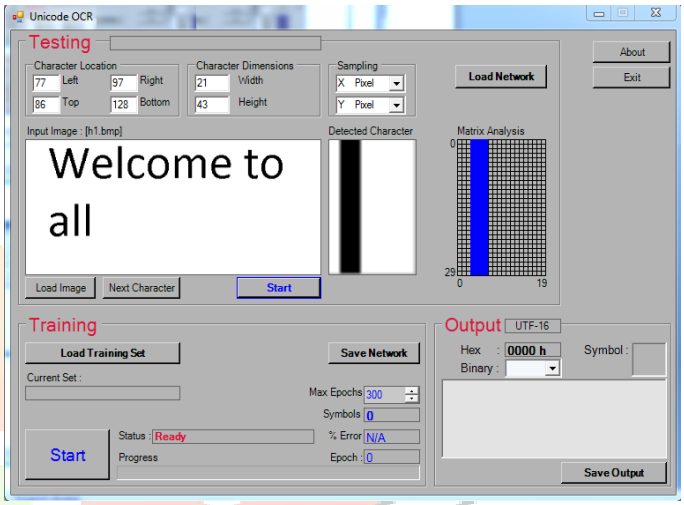
Number of characters=90, Learning rate=150, Sigmoid slope=0.014

Font Type	300		600		800	
	№ of wrong characters	% Error	№ of wrong characters	% Error	№ of wrong Characters	% Error
Latin Arial	4	4.44	3	3.33	1	1.11
Latin Tahoma	1	1.11	0	0	0	0









B. Results for variation in number of Input characters

Number of Characters=100, Learning rate=150, Sigmoid slope=0.014

Font Type	10		50		90	
	№ of wrong characters	% Error	№ of wrong characters	% Error	№ of wrong Characters	% Error
Latin Arial	0	0	6	12	11	12.22
Latin Tahoma	0	0	3	6	8	8.89
Latin Times Roman	0	0	2	4	9	10

C. Results for variation in Learning rate parameter

Number of characters=90, Number of Epochs=600, Sigmoid slope=0.014

Font Type	50		100		120	
	№ of wrong characters	% Error	№ of wrong characters	% Error	№ of wrong Characters	% Error
Latin Arial	82	91.11	18	20	3	3.33
Latin Tahoma	56	62.22	11	12.22	1	1.11
Latin Times Roman	77	85.56	15	16.67	0	0

5. Conclusion:

In this paper, proposed system to create a reliable OCR. This can be done by loading a whole word image, dividing it into letters using some clustering or/and an edge detecting techniques and sending the letter to my OCR with a flag indicating the letter location in the word (last letter or not). This can be done again by using clustering or/and a edge detecting techniques in order to find the location of the letter within the "image box" and moving it to the center.

Reference

[1] Y. Tang , L. Peng , Q. Xu, Y. Wang , and A. Furuhata ,"CNN Based Transfer Learning for Historical Chinese Character Recognition," in Proceedings of the 12th IAPR Workshop on Document Analysis Systems (DAS), 2018, pp. 25–29.

[2] Tan Chiang Wei , U. U. Sheikh Ab Al-Hadi Ab Rahman ,"Improved Optical Character Recognition with Deep Neural Network" 2018 IEEE 14th International Colloquium on Signal Processing & its Applications (CSPA 2018), 9 -10 March 2018, Penang, Malaysia.

[3] S.Thiyagarajan , Dr.G.Saravana Kumar , E.Praveen Kumar , G.Sakana ," Implementation of Optical Character Recognition Using Raspberry Pi for Visually Challenged Person", International Journal of Engineering & echnology, 7 (3.34) (2018) 65-67.

[4] Sumam Francis , Cannannore Nidhi Narayana Kamath , Andreas Dengel ," An Investigative Analysis of Different LSTM Libraries for Supervised and Unsupervised Architectures of OCR Training", 2018 16th International Journal on Frontiers in Handwriting Recognition.

[5] Martin Jenckel , Syed Saqib Bukhari , Andreas Dengel " Transcription Free LSTM OCR Model Evaluation "2018 16th International Journal on Frontiers in Handwriting Recognition

[6] G V S S K R Naganjaneyulu , A.V.Narasimhadhan , K Venkatesh ," Performance evaluation of OCR on poor resolution text document images using different pre processing steps",(2017).

[7] Mehmet Yasin AKPINAR, Erdem EMEKL˙IG˙IL, Seçil ARSLAN ,"Extracting Table Data from Images Using OpticalCharacter Recognition Text", (2017).

[8] Ms Poonam A. Wankhede , Dr. Sudhir W. Mohod ," A Different Image Content-based Retrievals using OCR Techniques", International Journal on Electronics, Communication and Aerospace Technology 2017

[9] Quang Anh BUI , David MOLLARD , Salvatore TABBONE ," Selecting automatically pre-processing methods to improve OCR performances ",2017 14th International Journal on Document Analysis and Recognition.

[10] Sebastien Eskenazi, Petra Gomez-Kr ´ amer, and Jean-Marc Ogier ," A study of the factors influencing OCR stability for hybrid security ",2017 14th International Journal Conference on Document Analysis and Recognition.

[11] Ido Kissos , Nachum Dershowitz ," OCR Error Correction Using Character Correction and Feature-Based Word Classification", 2016 12th Journal on Document Analysis Systems.

[12] Ayatullah Faruk Mollah , Nabamita Majumder , Subhadip Basu and Mita Nasipuri ,"Design of an Optical Character Recognition System for Camera-based Handheld Devices" ,IJCSI International Journal of ComputerScience Issues, Vol. 8, Issue 4, No 1, July 2015.

[13] Abdul Rahiman M† , Adheena C V , Anitha R , Deepa N , Manoj Kumar G , Rajasree M S ," Bilingual OCR System for Printed Documents in Malayalam and English" (2015).

[14] Peng Wan , Minoru Uehara ," Spam Detection Using Sobel Operators and OCR" ,2015 26th International Journal on Advanced Information Networking and Applications Workshops.

[15] M. Oquab, L. Bottou, I. Laptev and J. Sivic ,"Learning and Transferring Mid-level Image Representations Using Convolutional Neural Networks," in Proceedings of the 2016 Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1717–1724.

[16] Mande Shen , Hansheng Lei ," Improving OCR Performance with Background Image Elimination", 2015 12th International Journal on Fuzzy Systems and Knowledge Discovery (FSKD)

[17] Mohamed Fawzi1 , Mohsen. A. Rashwan 1 , Hany Ahmed 1 , shaimaa Samir 1 , SherifM. Abdou2 , Hassanin M. AI-Barhamtoshy3 , and Kamal M. Jambi3 ," Rectification of Camera Captured Document Images for Camera- Based OCR Technology", 2015 13th International Journal on Document Analysis and Recognition