A Review on Optical Character Recognition (OCR) & its Process

Honey Mehta
Department of Computer Science Engineering
IET Bhaddal, Ropar, Punjab,India

Sanjay Singla
Department of Computer Science Engineering
IET Bhaddal, Ropar, Punjab,India

Abstract:- Character recognition is a very complex task when one talking about its computerized implementation. This process mainly implemented in two different situations, One is when a document is provided as input to this system in the form of scanned image, the process is call offline character recognition and the other situation is when user write something in its own handwriting in parallel to the recognition work, this process is called Online recognition process. Offline form of recognition process is either used for handwritten characters or used for printed character recognition. In this paper we mainly focused on the step by step procedure of recognition process for printed characters belongs to roman script. Complexity of the procedure exists because of the variation on font styles used to print text, size of characters, fore color and background color in text etc.

Keywords –: OCR, Offline Recognition, Online recognition, Roman script.

I. INTRODUCTION

OCR is a process of conversion of data present in an image to its digital form for further use. The data given as input in case of OCR system is in the form of scanned images containing text portion [1,2]. The converted text by an OCR can be used for many other cases in which the text contained by image is failed to perform. The digital data can be edited, manipulated and stored easily than images, as images required large memory to store as compared to the digital data or the data present in computer's understandable form [3]. In case of its hardware and software requirements, OCR software is required during the process of conversion and a scanner is required as a hardware to input the image in to the system. In case of some application which require the storage of large amount of data to perform manipulation functions must use OCR to store and retrieve data in efficient way [4,5]. Rather to store data in the form of images. For example, users may require to scan and OCR a noteworthy newspaper or magazine article into a file that they can subsequently edit, save, print or e-mail to colleagues or business partners. Or they may require to incorporate some text into a presentation package, or convert it into HTML and post it up onto the Web site. For this purpose, the OCR activity is very simple: the user inserts a page with the text one wants captured into a scanner, the scanner makes an image of the text, the OCR program then processes the image, and translates and renders it into computer legible and editable text. OCR is a special case of *model-based* computer vision, it initially stores the prior information about the character set of any language, which it wants to convert to digital form later (in OCR, “characters”, with one model per character, for each font), and we are asked to all instances of those models(or approximations thereof) that appear in an image [6,7].

In case of OCR, an image is given as input to the system and it contains text in it [8]. Then the OCR system process the input image to extract some of the patterns present in the image for the purpose clarification. these patterns further used during the process of identification of characters present in the image [9,10].

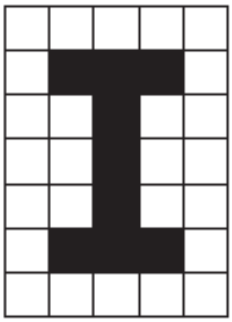

Figure 1: Text Image

Figure1. represents a text image in the form of a matrix where small blocks are there in white and black color. The black colored boxes represents the area in the image which is used to represent character and white porting is the part which is not participated in the formation of the character. The OCR software works in different parts as firstly it identifies the font of the character and the shape of the character present in the image against the information stored in its database. After the identification of single character OCR tries to identify the complete word present in the text by checking the combination of different characters to make meaningful word.

OCR software works by extracting different patterns and then use some classifiers such as Artificial Neural Network (ANN), Support Vector Machine (SVM), Principle Component Analysis (PCA) etc. These software are based on the rule based approach to identify the letters, characters and words. OCR works for the identification of symbols from the supplied images. These symbols may be printed character combinations or hand written character combinations. The main purpose is the recognition of character present in the image supplied to the system. The recognition process is further considered under two different categories. Either the process is offline recognition process of online recognition process. In case of offline recognition process. firstly either the data is written with the help of the handwriting of the user on the piece of the paper or present in printed form on the piece of the paper. Then this paper is scanned and the scanned copy is further provided as input image to the OCR system. But in case of Online process, the input is provided by the hand writing of the user only with the help of any digitizer. So in case of Online recognition both recognition process and input process works parallel to each other to produce results.

Next part of this paper will explain the character recognition process in detail for offline printed text recognition. After that few of the applications of OCR systems will discussed and then conclusion section will conclude the worth of the research.

II. CHARACTER RECOGNITION PROCESS

The purpose of OCR is already discussed, that it is used to convert the text data presented in the form of a paper either handwritten or printed to its digital form for further use. The data is firstly scanned and converted to an image form. Then this image is used as input for the OCR system. And finally OCR identifies and recognize the characters or text present in the image by using already stored knowledge.

The different phases of OCR process are explained as:

1. Image acquisition and Digitization
2. Pre-processing of input image
3. Segmentation for the ROI
4. Feature extraction
5. Classification
6. Post-processing

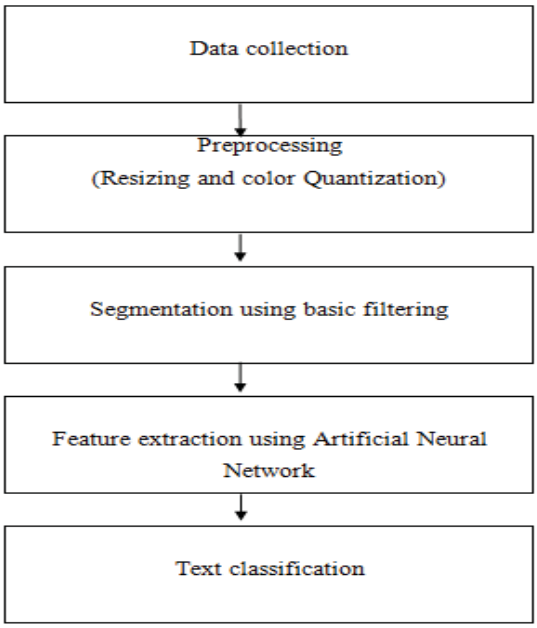

Figure 2. Character Recognition process

Image acquisition process involves in capturing the contents of the paper where the text is written [11]. Either that physical data is provided to the system by using the camera and clicking photographs. Or the second way is to use scanner for the purpose of input.

The next phase is the pre-processing of the image. Mainly this process can be done to remove the noise from the initial supplied image or to further enhance the product. This will provide an enhanced image to the further segmentation phase. Also as we discussed earlier image size is also an issue. So pre-processing step initially also perform resizing of image to handle the complexity of the process [12,13].

The process of segmentation is performed after pre-processing phase. The main idea behind the segmentation process is to extract some structure or simply a meaningful structure from the image by following the guidelines of continuity and discontinuity of image pixels [14]. Here in case of character recognition process, the segmentation process if followed to extract those areas from the image which participated in the formation of characters. Rest of the area is treated as the background for the image. In case of an image containing some text, the segmentation process first segmented the different lines containing text from the image data. After the extract of different lines, Word level segmentation is performed [15,16]. After segmentation of different words the segmentation is performed up to character level. Means from a single word different characters are identified which together formed that word

After segmentation phase, the next phase considered under the process of character recognition is feature extraction. In this phase the extracted components form the segmentation phase are analyzed further to find out a pattern which can uniquely identify the text. The main categories of features in character recognition process are: Statistical features and Structural features [17,18]. In case of statistical features the main things under consideration are curves, contour, Horizontal and Vertical projections etc. On the other hand structural features generally made estimation about the shape of the segmented portion. Structural features deal with loops, number of strokes, end points etc.

Classification process comes at the end when there is a need to make a decision about the presence of different characters as the text contained by an image as input to the OCR software [19,20]. Here the patterns are compared to the model set. The basis of this process of decision making is the features extracted under feature extraction process. The different approaches followed under this part are: Template matching, Statistical Classification, Structural classification and Artificial neural networks (ANN). In case of Template matching process, some character images are already stored in database and when an unseen text image is given as input to the system then similarity measure between the stored image and input image is used to identify the characters. During the process of structural classification, patterns are classified by their components and the relationship exists between the components. This process use decision trees and rule based systems. On the other hand an ordered and fixed length list of numerical data is used for the purpose of classification in case of statistical classification. Here the approach is followed as the mean value of the stored component is always compared to the actual contents of the feature identified. Another

method of classification is the use of Artificial neural network [21]. The different families of neural network like, Feed Forward Network, Back propagation network, Multi-layered perceptorn neural network etc. can be applied for the process of classification. During this process of ANN, a layered architecture is followed. Different layers of the network works for the purpose of classification. The easiest pattern in case of ANN is when it has only three layers to process. First is the input layer which can take the values for different neurons as weights from the segmented characters by the segmentation process. The second layer is training layer. This layer is used to compare the values of extracted pattern with some known contents stored by the neurons of middle layer of training layer. At the en Output layer is used to represent the classified characters.

Finally, the last stage is post-processing. It is used to improve the recognition result. This phase deals only with those symbols or characters which initially were failed to recognize by the recognition process. Dictionary look up is the main approach used in this phase for classification of characters or words. Here in this process the output generated by the OCR is compared with the different alternatives stored in some dictionary. These alternatives are mainly the predictions of characters and words. Then from these candidate solutions, the closest solution which matches with the OCR's output is selected as the best candidate solution and serves as a new output of the OCR software.

III. APPLICATIONS OF OCR

There are several benefits of in OCR software. The use of OCR will reduce the cost of the work in some business organizations. Some of these benefits listed as:

- *Storage Utilization:* To keep papers in an organization or important documents, there is a need of a lot of space. And this requirement increases with the growth of the organization. But if instead of storing records in different papers, scanned documents are preferred to be used and with the help of OCR systems these scanned documents can further be manipulated as per requirement.
- *Editable Text:* With the help of OCR the different generated documents are stored in text form and this text data further can be edited or modified as per the requirement. The files can further be converted into different forms like PDF, DOC etc.
- *Speed of Data Conversion and accuracy :* Data can also be converted to text form from scanned documents with the help of data entry operators through typing. But this is a very time consuming process as all depends on the speed of data entry operator. Another fault in the above approach is that the entry of erroneous data by the operators. So an efficient OCR system can handle these issues and provides an efficient and effective way of conversion of scanned documents to typed form.

IV. CONCLUSION

To understand the process of character recognition to an automatic system is a very complex task. This all happen because of the existence of different styles and size of fonts. The process involved in steps like pre-processing to enhance the input data, then Segmentation process involved to extract the desirable region of interest for further recognition, After that feature extraction process is used to search some specific contents for the classification of the character. This paper highlight the different processes which have to be perform in a sequence during recognition process and the benefits of OCR systems.

REFERENCES

[1] Srivastav,A., Kumar,J.,2010, “Text Detection in Scene Images using Stroke Width and Nearest-Neighbor Constraints”
[2] Lim,J., Kim,S., Park,J., Lee,G., Yang,H., Lee,C.,2009, “Recognition of Text in Wine Label images”
[3] Cui,J., Wang,L., Mei,J., Zhang,D., Wang,X., Peng,Y., Zhang,W.,2010, “CAPTCHA Design Based on Moving Object Recognition Problem”
[4] Mao,S., Huang,X., Wang,M.,2010,“An Adaptive Method for Chinese License Plate Location “
[5] Kumar,M., Kim,Y.C., Lee,G.S.,2010, “Text Detection using Multilayer Separation in Real Scene Images”
[6] Al-Shamma,S.D., Fathi,S.,2010,“Arabic Braille Recognition and Transcription into Text and Voice”
[7] Ikica,A., Peer,P.,2011,“An improved edge profile based method for text detection in images of natural scenes”
[8] Raj,S.B.E., Devassy,D., Jagannivas,J.,2011, “A New Architecture for the Generation of Picture Based CAPTCHA”
[9] Wazalwar,D., Oruklu,E., Saniie,J.,2011,“Design Flow for Robust License Plate Localization”
[10] Shivakumara,P., Phan,T.Q., Shijian Lu, Tan.C.L.,2011,“Video Character Recognition Through Hierarchical Classification”

[11] Chang,R.C.,2011,"Intelligent Text Detection and Extraction from Natural Scene Images"
[12] Zhang,S., Li,H.,2011, "Video Semantic Mining based on Dense Sub Graph Finding"
[13] Oi-Mean Foong, Hairuman,I.F.B., 2011, "OCR Signage Recognition with Skew & Slant Correction For Visually Impaired People"
[14] Wang,X., Lai,W., 2010, "Edge Detection for Chinese Text Image Based on Novel Differential Operator"
[15] Huang,X., 2012, "Automatic Video Text Detection and Localization Based on Coarseness Texture"
[16] Bo Bai, Fei Yin, Cheng-Lin Liu ,2012,"A Fast Stroke-Based Method for Text Detection in Video"
[17] Shih-Chung Chen, Chung-Min Wu, Shih-Bin Su, 2012,"Image Morse Code Text Input System"
[18] Mello,C.A.B., Costa, D.C., Santos,T.J.D., 2012, "Automatic Image Segmentation of Old Topographic Maps and Floor Plans"
[19] Zhang,Y., Wang,C., Xiao,B., Shi,C., 2012, "A New Text Extraction Method Incorporating Local Information"
[20] Yang,Z., Shi,P.,2012,"Caption Detection and Text Recognition in News Video"
[21] Rajeshbaba, M., Anitha, T., 2013, "Detect and Separate localization Text in Various Complicated-Colour Image"