Optical Character Recognition Systems- A Survey

Shruti Ahuja
Assistant Professor, D. A. V College, Abohar, Punjab, India.

Received: July 14, 2018 Accepted: August 25, 2018

ABSTRACT *Optical Character Recognition (OCR) has been a topic of interest for the researchers in last few decades. It involves digitizing a document image into its component characters. Despite decades of research, developing OCR with strong capabilities comparable to that of humans is still considered to be a challenging job. Due to this challenging nature, academic and industrial researchershave directed their attentions towards Optical Character Recognition. Over the last few years, large number of academic laboratories and companies got involved in research on Character Recognition. This research aims at summarizing the research done in the field of OCR and presents an overview of different aspects of OCR, issues faced and proposals to resolve the issues of OCR.*

Keywords:

1. INTRODUCTION

Optical Character Recognition (OCR) is a software that converts printed text into digitized form such that it can be manipulated by machine. Unlike human brain, machines are not intelligent enough to perceive the information obtained by an image. Therefore, a lot of research efforts have been put forward to transform a document image into a format understandable by the machine. OCR is a complex problem because ofdiversity of languages, fonts and styles of text and the multiple rules of languages. Hence, techniques from different disciplines of computer science (i.e. image processing, pattern classification and natural language processing) are used to address different challenges in character recognition process. This paper familiarizes the reader with the problem and enlightens the reader with the historical perspectives, challenges, techniques and applications of OCR.

2. TYPES OF OPTICAL CHARACTER RECOGNITION SYSTEMS

There has been multiple objectives for which research on OCR has been carried out during the past years. This section explains different types of OCR systems that have emerged as a result of these researches. Character recognition systems are used to encode the scanned images of documents which are hand written, type written or printed into machine understandable text. This translated machine encoded script can be easily edited and searched. The translated text can also be processed in various other forms according to the requirements. The storage requirement for the scanned documents is also less as compared to the requirement for whole scripts. Manual handling and processing of the written scripts is complex task, whereas character recognition systems have reduced the human chore to a greater extent. For the recognition of individual characters from the entire documents, computerized processing is required. We can classify character recognition systems on the basis of mode of image acquisition, character connectivity, font-restrictions etc. Fig. 1 shows the classification of character recognition systems. On the basis of type of input, the OCR systems can be categorized as handwriting recognition and machine printed character recognition.

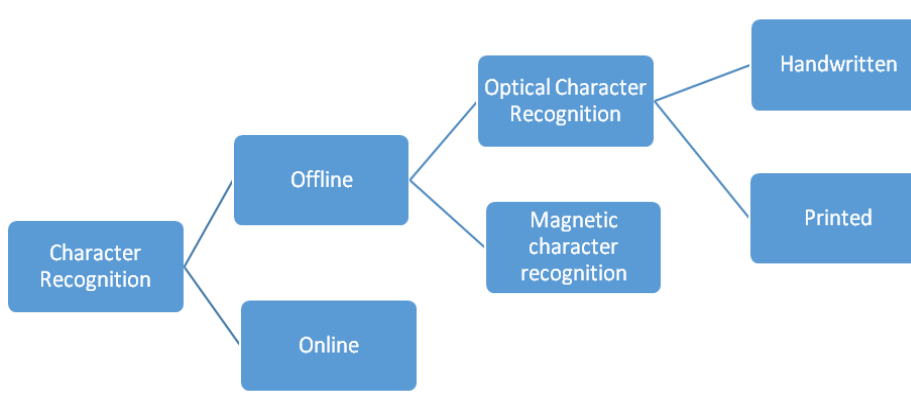

Fig. 1: Classification of Character Recognition System

Character recognition can be performed in two forms online and offline. Online character recognition is commonly used to recognize handwritten characters.It involves capturing the data during the writing process using a special pen on an electronic surface. As the pen moves on that electronic surface, the coordinates of successive points are represented as a function of time. In recent times, due to amplified use of handheld devices, online handwritten recognition have grabbed the attention of researchers. Onlinerecognition systems targets to provide a natural interface to the users to type on screen by handwriting using a pad instead of typing by the keyboard. In online handwriting recognition, it is very easy for the users to detect and correct the wrongly recognized characters on the spot. This can be performed by verifying the recognition results as soon as they appear. The user must modify his writing style so as to improve the recognition accuracy. Also, a machine can be trained to recognizeparticular user's writing style. So both machine adaptation and writer adaptation is possible.There have been many online systems available because they are easier to develop, have good accuracy and can be incorporated for inputs in tablets and PDAs [1].

While in the offline mode of character recognition, prewritten data written on a sheet of paper is scanned. So all type-written or printed characters are classified in offline mode. Off-line handwritten character recognition involves recognizing the characters in a given document that have been scanned from a surface and are stored digitally in gray scale format. The scanned documents needshuge amount of storage and also require several processing applications such assearching for a content, editing and maintenance. Such documents require humans to processthem manually. Character recognition systems are required to translate scanned imagesof printed, typewritten or handwritten text into machineencoded text.

3. MAJOR PHASES OF OPTICAL CHARACTER READER

The process of OCR is a composite activity comprises different phases starting from image acquisition to classification. Character recognition system comprises ofvarious steps to completely recognize and encode the text into machine text. The image of the document to be read is captured from an external source such as scanner or camera.Several phases involved in character recognition after acquisition of document image are: Pre-processing, Segmentation, Feature extraction and Classification. The proposed recognition system is shown using a block diagram in Fig. 2.

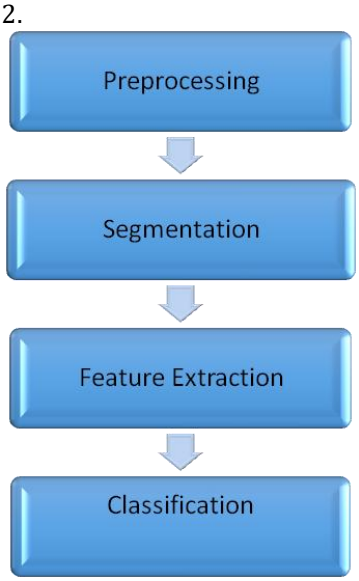

Fig.2: Block Diagram of character Recognition system

a) Pre-Processing

In pre-processing phase, there is a sequence of operations that are performed on the scanned image. It improves the image rendering making it suitable for segmentation of grayscale character image into a window sized. The scanned image may have some noise due to some unnecessary detailspresent in the image.After noise reduction, the image is then saved for later use. All the pre- designed templates of the alphabets are loaded into the system.One of the techniques used for pre-processing is thresholding that converts the image into binary formatbased on some threshold value [2]. That threshold value can be set at local or global level. Different types of filters such as averaging, max and min filters can be used. One more important part of pre-processing includes finding out the skew in the document. Different techniques used for skew estimation are: projection profiles, Hough transform, nearest neighbourhood methods etc. In some

cases, thinning of the image is also performed before later phases are applied [3]. Finally, the text lines present in the document can also be found out as based on clustering or projections of the pixels.

b) Segmentation

In segmentation, the position of the character in the image is found out and the size of the image is converted to that of the template size. Segmentation of handwritten characters is more difficult than that of printed documents in the standard form. This is chiefly because of the variability in paragraph, words of a line and characters of word. The characters can further be slant, curved or skewed.

c) Feature Extraction

Features of individual characters are extracted in this phase. The performance of each character recognition system depends on the features extracted. The extracted features from an input character should allow classification of each character in an inimitable way. The selection of the accurate features and number of features to be used is an important research question.Various features that can be usedto find the feature set for a character are diagonal features, transition features, intersection and open end points features, zoning features, directional features, power curve fitting–based features and parabola curve fitting–based features.Also, several techniques such as principal component analysis can be used to reduce the dimensionality of the image.

d) Classification

In classification, the test object is compared with the system database containing predefined sample of objects to classify its appropriate class. In supervised classification, a group of known pixels are used to train a classifier. The trained classifier is further used to classify other images.The structural approach to classification is based on relationships in the image components. The statistical approaches use a discriminate function to classify the image. Some of the statistical classification approaches are Bayesian classifier, decision tree classifier, neural network classifier, nearest neighbourhood classifiers etc. [4]. Finally, there are classifiers based on syntactic approach that assumes a grammatical approach to derive an image from its sub- constituents.

4. IMPROVING ACCURACY OF OCR SYSTEMS

After classification, the results of character recognition are not 100% correct, especially for complex languages. Post processing techniques need to be performed to improve the accuracy of OCR systems. These techniques use natural language processing, geometric and linguistic framework to suggest and correct errors in OCR results. For example, post processor can employ a spell checker and dictionary, probabilistic models like Markov chains and n-grams to improve the accuracy. The time and space complexity of a post processor should not be high and the application of a post-processor should not cause new errors. To improve the accuracy of OCR results, we can use more than one classifier for classification. Multiple classifiers can be used in parallel, cascading, or hierarchical fashion. The results of different classifiers can then be combined. Also, to improve OCR results, contextual analysis can be done. The geometrical and document context of the image can help in reducing the chances of errors. Lexical processing based on Markov models and dictionary can also help in improving the results of OCR [4].

5. APPLICATIONS OF OCR

OCR supports a large number of useful applications. In the early days, OCR has been used for sorting the mails, reading bank cheques and verifying signatures [5]. Further, OCR has been used by various for automated form processing in places where anenormousdata is available in printed form. Another valuable application of OCR is in helping blind and visually impaired people to read text [6]. Other uses of OCR include passport validation, processing utility bills, pen computing and automated number plate recognition [7].

6. CONCLUSION

In this paper, an outline of various techniques of OCR has been presented. OCR comprises of various phases such as acquisition, pre- processing, segmentation, feature extraction and classification. Each of the steps is discussed in this paper. In future, work can be done to develop an efficient OCR system by combining multiple different techniques. The OCR system can also be used in different practical applications such as smart libraries, number-plate recognition and various other real-time applications. Although significant amount of research in OCR have been done, recognition of characters for language such as Arabic, Sindhi and Urdu is still considered a challenge. Another important area which needs advance research is multi-lingual character recognition system. Finally, various applications of OCR systems in practical situations is an active area of research.

REFERENCES

1. Qadri, M.T., & Asif, M, 2009, Automatic Number Plate Recognition System for Vehicle Identification Using Optical Character Recognition presented at International Conference on Education Technology and Computer, Singapore, 2009. Singapore: IEEE.
2. Lund, W.B., Kennard, D.J., & Ringger, E.K. (2013). Combining Multiple Thresholding Binarization Values to Improve OCR Output presented in Document Recognition and Retrieval XX Conference 2013, California, USA, 2013. USA: SPIE
3. Shaikh, N.A., & Shaikh, Z.A, 2005, A generalized thinning algorithm for cursive and non-cursive language scripts presented in 9th International Multitopic Conference IEEE INMIC, Pakistan, 2005. Pakistan: IEEE
4. Ciresan, D.C., Meier, U., Gambardella, L.M., & Schmidhuber, J, 2011, Convolutional neural network committees for handwritten character classification presented in International Conference on Document Analysis and Recognition, Beijing, China, 2011. USA: IEEE.
5. Shen, H., & Coughlan, J.M, 2012, Towards a Real Time System for Finding and Reading Signs for Visually Impaired Users. Computers Helping People with Special Needs. Linz, Austria: Springer International Publishing.
6. Bhammar, M.B., & Mehta, K.A, 2012, Survey of various image compression techniques. International Journal on Darshan Institute of Engineering Research & Emerging Technologies, 1(1), 85-90.
7. Bhavani, S., & Thanushkodi, K, 2010, A Survey On Coding Algorithms In Medical Image Compression. International Journal on Computer Science and Engineering, 2(5), 1429-1434.