Design of Machines and Structures, Vol. 7, No. 2 (2017), pp. 5–12.

A REVIEW OF OPTICAL CHARACTER RECOGNITION SYSTEM

HMOUMEN MAROUANE

University of Miskolc, Robert Bosch Department of Mechatronics
3515 Miskolc-Egyetemváros
hhmoumen.marouane@gmail.com

Abstract: Optical character recognition (OCR) has been a topic of great interest for many years. It is a system that permits us to convert various types of documents into machine encoded/computer-readable text. It consist of steps like image acquisition, pre-processing, segmentation, feature extraction, etc. The purpose of this work is to summarize the researches performed in the OCR field. It provides an overview of different aspects of OCR and discusses corresponding proposals aimed at resolving problems of OCR. A practical OCR problem is also investigated.

Keywords: *OCR, image acquisition, pre-processing, segmentation, feature extraction*

1. INTRODUCTION

Optical character recognition (OCR) is a system that allows us to convert various types of documents (PDF, BMP, TIFF, JPEG, PNG) into machine computer-readable text. It has become one of the most outstanding applications of technology in the domain of artificial intelligence and pattern recognition. Contrarily to the human brain which has the power to recognize easily the characters/text from an image, machines are still far to reach the human level to perceive the information available in image. Consequently a large number of research efforts have been put forward that attempts to convert efficiently a document image to format understandable for machine.

OCR is a sophisticated problem because there is a lot of variables that can affect the detection of the text/characters such the diversity of the languages, styles, and fonts in which text can be written also the environmental light that is difficult to control, etc. Therefore, techniques from different disciplines of computer science are employed to address different challenges.

This paper is organized as follows. In section 2, the different types of optical recognition systems will be studied. The components of the OCR will be shown in section 3. A practical problem is analyzed in section 4. Finally, some conclusions are given in the last section.

2. TYPES OF OPTICAL CHARACTER RECOGNITION SYSTEMS

There has been plenty of directions in which investigation on OCR has been achieved. This part review different types of OCR systems that have appeared as outcome of these researches. We can classify these systems based on character connectivity, font-restrictions, image acquisition mode, etc. *Figure 1* classifies the character recognition systems.

According to the type of the input, the OCR system can be classify as machine printed character recognition or handwriting recognition. The former is relatively simpler problem because characters are usually of uniform dimensions, and the positions of characters on the page can be predicted [1].

Handwriting recognition is arranged into two types as on-line and off-line character recognition. Off-line handwriting recognition includes automatic conversion of text into an image into letter codes which are applicable within computer and text-processing applications. Off-line handwriting recognition is harder, as a lot of people have different handwriting styles. But, in the on-line system, on-line character recognition deals with a data stream which comes from a transducer while the user is writing. The typical hardware to collect data is a digitizing tablet which is electromagnetic or pressure sensitive. When the user writes on the tablet, the successive movements of the pen are transformed to a series of electronic signals which is memorized and analyzed by the computer.

There have been numerous online systems usable because they are easy to develop, have good accuracy and can be integrated for inputs in tablets [2].

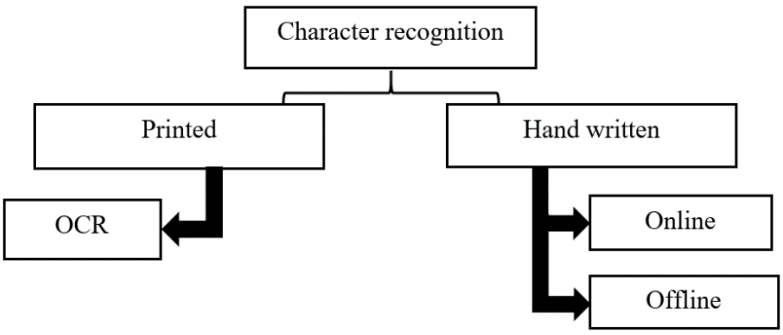

Figure 1. Types of character recognition system

3. COMPONENTS OF AN OCR SYSTEM

The principal notion in automatic recognition of patterns is first to train the machine which class of patterns that can appear and what they look like [3, 4]. In OCR patterns are numbers, letters and several special symbols like slash, exclamation, etc. The machine training is achieved by displaying to the machine examples of characters of all different classes. Based on these examples the machine builds prototype

or description of each class of characters. Throughout recognition the unknown characters are matched to previously obtained descriptions and assigned to class that gives the best match.

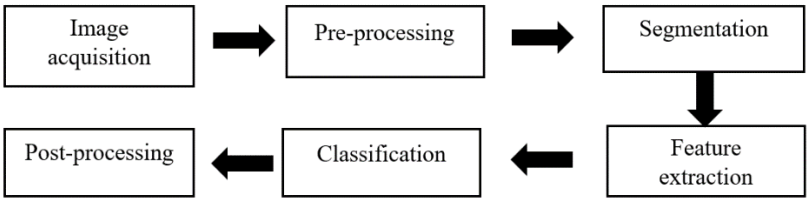

Figure 2. Components of an OCR-system

An OCR system is composed of many components as exposed in *Figure 2* [3, 5]. The initial step is to digitize analog document using an optical scanner. When regions consisting text are discovered each character is extracted by segmentation process. The obtained characters are pre-processed, removing noise to simplify feature extraction. The identity of each character is detected by matching extracted features with descriptions of character classes obtained by a previous learning phase. Lastly contextual data is used to rebuild numbers and words of the initial text. These steps are shortly given in *Figure 2*.

3.1. Image acquisition

Image acquisition is the first step of OCR system that involves getting a digital image from an external source like scanner or camera and converting it into appropriate form that can be easily processed by computer. This can include compression as well as quantization of image. A special case of quantization is binarization it converts an image of up to 256 gray levels to a black and white image. Generally, the binary image is sufficient to characterize the image. An overview of various image compression techniques have been provided in [6].

3.2. Pre-processing

Once the image has been acquired, different preprocessing steps can be performed to improve the quality of image and to make it suitable for recognition. A typical OCR system may use the following techniques for image enhancement:

- Filtering: The aim of it is to remove noise and diminish spurious points usually introduced by poor sampling rate of the data acquisition device and uneven writing surface. The main idea is to convolute a predefined mask with the image to assign a value to a pixel as a function of the gray values of its neighboring pixels. Various filters have been created for thresholding, smoothing, removing lightly colored background and contrast adjustment purposes [7, 8].
- Skewed correction: The camera captures images may suffer from skew and perspective distortions. This effect is due to improper image capture technique

like the angle of the camera with the object or the lens of the camera. The horizontal text may suffer rotation at some degrees. The calculation method for rotation of the image has been described in reference [9].
- Thinning: Is the process of reducing the width of the foreground pixels. While thinning, it is necessary to maintain the form of the characters on the image. Thinning is done on the basis of neighborhood of a pixel, e. g., if a line on an image is of 3 pixels width, the thinning function will change the border pixels of the line and the output image will consist of line of one pixel width.

3.3. Segmentation

Segmentation is a process that defines the components of an image, and its task is to isolate the characters or words from the back ground of the picture or the document. The segmentation can be done explicitly or implicitly as a byproduct of classification phase [10]. Most of optical character recognition algorithms segment the words into isolated characters which are recognized separately. Normally this segmentation is achieved by isolating each connected component that is each connected black area. This technique is easy to implement, but problems appear if characters touch each other or if characters are fragmented and consist of various parts.

3.4. Feature extraction

The aim of feature extraction is to catch the main characteristics of the characters/symbols, and it is usually admitted as one of the most difficult problems of pattern recognition. The right way of describing a character is to use the actual raster image. Another approach is to extract some features that still characterize the characters/symbols but without taking into consideration the unimportant attributes. The technics for extraction of such features are usually split into three principal groups where the features are found from:
- Structural analysis.
- The distribution of points.
- The distribution of points.

3.5. Classification

It is described as the process of classifying a character into its correct class. The basic approach to classification is founded on links present in image components. The analytical approaches are based on the usage of a discriminate function to arrange the image. Some of the statistical classification approaches are Decision tree classifier, Bayesian classifier, Neural network classifier, etc. Finally, there are classifiers based on syntactic approach that assumes a grammatical approach to compose an image from its sub-constituents.

3.6. Post-processing

After the classification of the character, there are several approaches that can be adopted to enhance the preciseness of the OCR results. One of the paths is to utilize several classifier for the classification of the image. The classifier can be used in parallel, hierarchical or cascading fashion. The results of the classifiers can then be united using several approaches. For a better OCR results, contextual analysis can also be executed. The geometrical and document context of the image can aid in decreasing the chances of errors. Lexical processing based on Markov models and dictionary can also help in improving the results of OCR [11].

4. A PRACTICAL PROBLEM

In this part, we will show you an experiment that have been performed using iRvision system that is a ready-to-use robotic vision system which requires no additional hardware except for a camera or sensor and cable. It provides a 2-D or 3-D robot guidance and/or error proofing tool to accomplish part location, presence detection, and other operations that normally require special sensors or custom fixturing. In our case the experiment will use the iRvision system to recognize successfully scratched numbers. It should be mentioned that in the teaching process of the system a non-scratched numbers were used. The laboratory is located in University of Miskolc, at Robert Bosch Department of Mechatronics. iRVision consists of the following components (see in *Figure 3*):

- Camera and lens, or three-dimensional laser sensor.
- Camera cable.
- Optional multiplexer (contained in the robot controller).
- Setup PC ... *.
- Communication cable ... *.

Note: The components marked with an asterisk (*) are necessary only for setting up iRVision and can be removed during production operation.

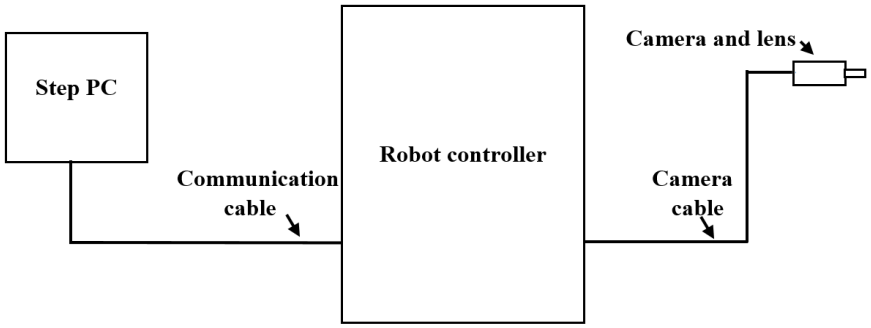

Figure 3. The iRVision System

After teaching our iRvision system to recognize number 3, we created some scratched number that was written in a piece of paper A4 and this scratches varied from big to small one. Then iRvision system was applied to verify if it can recognize the scratched numbers successfully in such cases when changing some parameters like score threshold, contrast threshold, scale, orientation. Consequently these parameters increase the necessary time for recognition. The application of the solution is shown in the *Figure 4*. Environment light was used in this experiment.

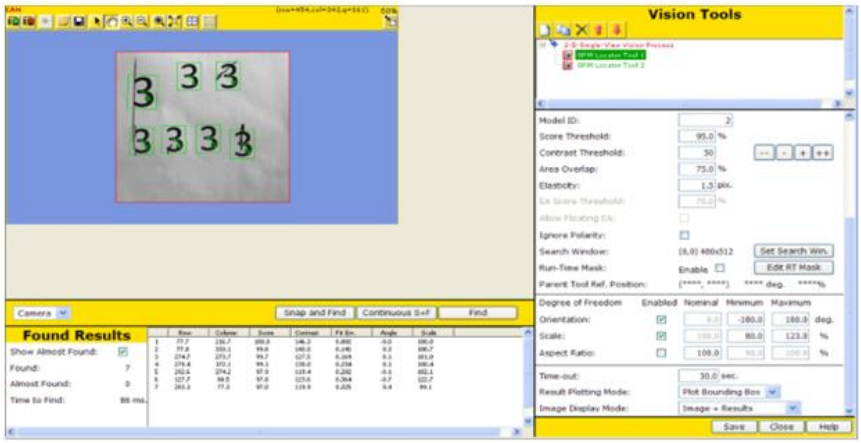

Figure 4. Results of the iRvision system

All the scratched numbers were detected successfully but in the Found Results section there are some parameters which are different from each other:

- Row and Column: Coordinate values of the model origin of the found pattern (units: pixels).
- Score: Represent how similar the pattern found in the image is to the model pattern. It is ranging from 0 to 100 points. If the pattern fully matches, it gets a score of 100 points. If it does not match at all, the score is 0.
- Contrast: This value represents "how clearly the pattern found in the image can be seen". The value of contrast ranges from 1 to 255. The larger the value, the clearer the pattern.
- Angle and scale of the found pattern: The Angle of the found pattern indicates the degree of rotation with respect to the model pattern. The scale of the found pattern shows the value of how many times it is expanded with respect to the model pattern.
- Fit error (Fit err.): Deviation of the found pattern from the model pattern (units: pixels).

There is a lot of system which performs better like Cognex system. Using thus system allows us to apply morphological operation to remove noises from the image and make it easy to be detected.

5. SUMMARY

An overview of several techniques of OCR has been discussed in this paper. An OCR is not a single process, however it includes several stage like acquisition, preprocessing, segmentation, feature extraction, classification and post-processing. Each steps is reviewed in this paper. Using a combination of these techniques, a powerful OCR system can be developed as a future work. The OCR system can also be used in several functional applications such as smart libraries, number-plate recognition and various other real-time applications used in the industry.

6. ACKNOWLEDGEMENT

The described article/presentation/study was carried out as part of the EFOP-3.6.1-16-2016-00011 "Younger and Renewing University – Innovative Knowledge City – institutional development of the University of Miskolc aiming at intelligent specialization" project implemented in the framework of the Szechenyi 2020 program. The realization of this project is supported by the European Union, co-financed by the European Social Fund.

REFERENCES

[1] BHANSALI, M.–KUMAR, P.: An Alternative Method for Facilitating Cheque Clearance Using Smart Phones Application. *International Journal of Application or Innovation in Engineering & Management (IJAIEM)*, 2013, 2 (1), 211–217.

[2] QADRI, M. T.–ASIF, M.: *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.

[3] CHAUDHURI, A.: *Some Experiments on Optical Character Recognition Systems for different Languages using Soft Computing Techniques*. Technical Report, Birla Institute of Technology Mesra, Patna Campus, India, 2010.

[4] CHERIET, M.–KHARMA, N.–LIU, C. L.–SUEN, C. Y.: *Character Recognition Systems: A Guide for Students and Practitioners*. John Wiley and Sons, 2007.

[5] RICE, S. V.–NAGY, G.–NARTKER, T. A.: *Optical Character Recognition: An Illustrated Guide to the Frontier*. The Springer International Series in Engineering and Computer Science, Springer US, 1999.

[6] LUND, W. B.–KENNARD, D. J.–RINGGER, E. K.: *Combining Multiple Thresholding Binarization Values to Improve OCR Output* presented in Document Recognition and Retrieval XX Conference 2013, California, USA, 2013. USA, SPIE.

[7] ARICA, N.–VURAL, F. T. Y.: An Overview of Character Recognition focused on Offline Handwriting, IEEE Transactions on Systems, Man and Cybernetics – Part C. *Applications and Reviews*, 2001, 31 (2), 216–233.

[8] CHAUDHURI, A.: *Some Experiments on Optical Character Recognition Systems for different Languages using Soft Computing Techniques*. Technical Report, Birla Institute of Technology Mesra, Patna Campus, India, 2010.

[9] MOLLAH, A. F.–BASU, S.–DAS, N.–SARKAR, R.–NASIPURI, M.–KUNDU, M.: Text/Graphics Separation and Skew Correction of Text Regions of Business Card Images for Mobile Devices. *Journal of Computing*, Vol. 2, Issue 2, February 2010.

[10] SHAIKH, N. A.–SHAIKH, Z. A.–ALI, G.: *Segmentation of Arabic text into characters for recognition* presented in International Multi Topic Conference. IMTIC, Jamshoro, Pakistan, 2008. Pakistan, Springer.

[11] CIRESAN, D. C.–MEIER, U.–GAMBARDELLA, L. M.–SCHMIDHUBER, J.: *Convolutional neural network committees for handwritten character classification* presented in International Conference on Document Analysis and Recognition, Beijing, China, 2011, USA, IEEE.