International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 2 Issue 12, December - 2013

A Review of Optical Character Recognition

Jagruti Chandarana*, Mayank Kapadia**

* Department of Electronics and Communication Engineering, UKA TARSADIA University.

**Assistant Professor,Department of Electronics and Communication Engineering, UKA TARSADIA University.

Abstract

Optical Character Recognition, popularly referred to as the OCR, is an active field of research which has enormous scientific and practical interest and it is a special case of the Pattern Recognition Systems. Input characters are first digitized by an optical scanner in OCR systems. Each character is then located and segmented, and the resulting character image is fed into a preprocessor for noise reduction and normalization. Certain characteristics are then extracted from the character for classification. The feature extraction is a crucial step and many different techniques exist, each having its importance. This paper presents an overview of feature extraction methods for character recognition. Feature extraction method selection is the only most important factor in achieving high recognition performance in character recognition systems. For different representations of the characters different feature extraction methods are designed. When a few promising feature extraction methods have been identified, they need to be evaluated experimentally to find the best method for the given application.

Index Terms- Optical character recognition**,** Feature extractionmethods, Recognition methods,Performance evaluation measures.

1. INTRODUCTION

Pattern recognition is the assignment of a physical object or event to one of several pre-specified categories [1]. It has many applications such as"classification and analysis of RADAR signaling, character (letter or number) recognition, and handwriting analysis ('notepad' computers)". Other applications include bank checks, tablet computers, personal digital assistants (PDAs), Cheque reading, postcode recognition, form processing, and signature verification [2]. Optical character recognition has many different practical applications. The main areas where OCR has been of importance are text entry (office automation), data entry (banking environment) and process automation (mail sorting) [20].

Character is the basic building block of any language that is used to build different structure of a language**.** Characters are the alphabets and the structures are the words, strings and sentences etc. [7]. Character recognition techniques as a subset of pattern recognition give a specific symbolic identity to an offline printed or written image of acharacter [8]. Character recognition is better known as optical character recognition because it deals with therecognition of optically processed charactersratherthan magnetically processed ones. The main objective of character recognition is to interpret input as a sequence of characters from an already existing set of characters. The advantages of the character recognition process are that it can save both time and effort. It provides a fast and reliable alternative to typing manually.

Recognition of any character is a process which loads that character image, preprocesses the image, extracts proper image features, classify the characters based on the extracted image features and the known features are stored in the vectors, and recognizes the image according to the degree of similarity between the loaded image and the image databases.

2. Methods of Optical Character Recognition

There are various methods of the character recognition which can be divided into the following groups:

- Pattern systems;
- Structural systems;
- Feature systems;
- Neural network systems

The above mentioned systems have both advantages and disadvantages which are as follows:

1. Structural algorithms are very sensitive to the image defects. Besides, in opposition to the pattern and feature systems, effective automated learn procedures for structural systems are not implemented yet [3, 4, 5]

2. Feature systems loose important information while calculating the character features and as a consequence make errors on objects classification referring them to the wrong classes [3].

3. Although neural networks are able to recognize different fonts taking into consideration their defects and distortions, nevertheless they require complicated multi-layer structure and need a long training using sets of samples [6]. This is not always practicable in industrial environment and at the same time the economic forces are of great importance here.

4. Pattern algorithms are stable to small defects of the image and have sufficiently high recognition velocity. However even minor distortions of the image, which lead to the characters distortion, may influence negatively on the result of recognition [3, 4, 5]. The task which is to be solved by the certain algorithm has fast, reliable and stable recognition,i.e. when the probability of acquiring distorted and noised images is very high. Therefore appropriate recognition algorithm should meet the following requirements:

• Stability to defects of the recognized characters,
• High velocity,
• Easiness of the tuning and training.

When considering the existing methods of the optical character recognition and the mentioned above requirements, it was concluded that pattern algorithms are the most suitable for solving the appropriate task.

3. Methodologiesof OCR Systems

The character recognition system involves many steps to completely recognize and produce machine encoded text. In this section, we focus on the methodologies of CR systems. The literature review in the field of CR involved various phases are termed as: Pre-processing, Segmentation, Feature extraction and Recognition and Classification. The block diagram of character recognition system is shown in figure 1.

3.1. Pre-processing

Preprocessing plays an important role in an OCR system. In this method preprocessing consists of: binarization and size normalization which are explained as follows.

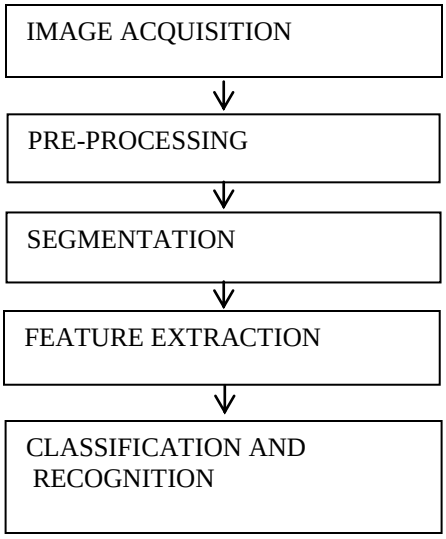

Figure 1: Block diagram of the character recognition system

3.1.1. Binarization

In order to reduce storage requirements and to increase processing speed, gray-scale or color images are often desirable to represent as binary images. This process is called Digitization of image.

3.1.2. Size Normalization

To account for variety in size, the character is normalized. By normalization, the character is made to fit into a standard size array. This size of array is chosen by trial and error method & the value that gives the best results is fixed. Any size of characters and shape can be processed and matched with the normalization technique.

3.2. Segmentation

In segmentation, the position of the character in the image is found out and the size of the image is normalized to that of the template size. Segmentation can be external and internal. External segmentation is the isolation of various writing units, such as paragraphs, sentences or words. In internal segmentation an image of sequence of characters is decomposed into sub-images of individual character.

3.3. Feature Extraction Methods

In most of the recognition systems, to avoid extra complexity and to increase the accuracy of the algorithms, a more compact and characteristic representation is required. For the sake of simplicity, the extraction of a set of features for each class that helps to differentiate it from other classes while remaining invariant to characteristic differences within the class. There are several feature extraction techniques given as follows

3.3.1. **Template Matching**.Matching covers the groups of techniques based on similarity measures where the distance between the feature vectors, describing the extracted character and the description of each class is calculated. Different measurement criteria may be used, but the common is the Euclidean distance. This minimum distance classifier works better when the classes are well separated, that is when the distance between the means is large compared to the spread of each class. When the entire character is used as input to the classification, and no features are extracted (template-matching), a correlation approach is used. Here the distance between the character image and database images representing each character class is computed.

3.3.2. Zoning.The frame containing the character is divided into several overlapping or non-overlapping zones. The densities of the points or some features in different regions are analyzed and form the representation [14]. For example, contour direction features measure the direction of the contour of the character [9, 10] which is generated by dividing the image array into rectangular and diagonal zones and computing histograms of chain codes in these zones.

3.3.3. Moments.The word 'moment' here refers to the some of the characteristics that can be calculated from the images. There are moments of different orders that are used in pattern recognition as they are in statistics.
In this case the moments of different points present in a character are utilized as a feature. These are most commonly used methods in character recognition. Moments, such as central moments and Zernike moments, form a compact representation of theoriginal document image that make the process of recognizing an objectscale, translation, and rotation invariant [12], [13]. Moments are considered as series expansion representation since the original imagecan be completely reconstructed from the moment coefficients. Hu's other moments are statistical measure of the pixel distributionabout the Centre of gravity of the character.

3.3.4. **Chain coding.**For detecting loops and curves in a character 'Freeman chain coding' is suggested in [15]. Chain codes are used to represent a boundary by a connected sequence of straight line segments of specified length and direction. The direction of each segment is coded by using a numbering scheme.

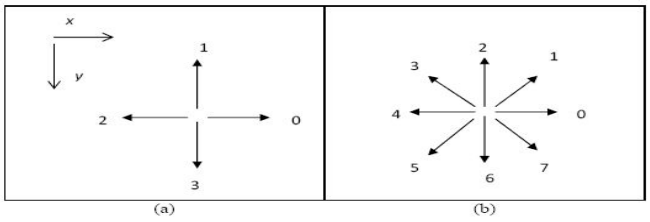

Figure 2: Direction numbers for (a) 4-directional chain code(b) 8-directional chain code.

A chain code can be generated by following a boundary of an object in a clockwise direction and assigning a direction to the segments connecting every pair of pixels. Freeman's chain coding is essentially obtained by mapping the strokes of a character into a 2-D parameter space, which is made up of codes, as shown in Fig. 2.

3.4. Recognition and Classification Techniques

3.4.1. Neural Networks (NNs). An NN is defined as a computing architecture that consists of a massively parallel interconnection of adaptive "neural" processors. Because of its parallel nature, it can perform computations at a higher rate compared to the classical techniques. Because of its adaptive nature, it can adapt to changes in the data and learn the characteristics of input signal. An NN contains many nodes. The output from one node is fed to another one in the network and the final decision depends on the complex interaction of all nodes. In spite of the different underlying principles, it can be shown that most of the NN architectures are equivalent to statistical pattern recognition methods [16]. The most common NNs used in the CR systems are the multilayer perceptron of the feedforward networks and the Kohonen's self-organizing map (SOM) of the feedback networks. Multilayer perceptron proposed by Rosenblatt [17] and is applied in CR.
Recently, the use of neural networks to recognize characters (and other types of patterns) has resurfaced. When a back-propagation network is considered, this network is composed of several layers of interconnected neurons. At the input layer a feature vector enters in the network. Each element of the layer computes a weighted sum of its input and transforms it into an output by a nonlinear function [19]. During training the weights at each connection areadjusted until a desired output is obtained. In OCR themain problem of neural networks may be their limited predictability and generality, and advantage is their adaptive nature.

3.4.2.Optimum statistical classifiers. Most important Kernel methods are Support Vector Machines,Principal Component Analysis (PCA),and Kernel Principal Component Analysis (KPCA) etc. Support vector machines (SVM) are a group of supervised learning methods that can be applied to classification. In a task of Classification usually data is divided into training and testing sets. The aim of SVM is to produce a model, which predicts the target values of the test data. Different types of kernel functions of SVM are: Linear kernel, Polynomial kernel, Gaussian Radial Basis Function (RBF) and Sigmoid.

3.5. OCR Performance Evaluation

No standardized test sets exist for character recognition, and as the performance of an OCR system is highly dependent on the quality of the input, this makes it difficult to evaluate and compare different systems. Still, recognition rates are often

given, and usually presented as the percentage of characters correctly classified. However, this does not say anything about the errors committed. Therefore in evaluation of OCR system, three different performance rates should be investigated:

3.5.1. Recognition rate.

The proportion of correctly classified characters.

3.5.2. Rejection rate.

The proportion of characters which the system was unable to recognize. Rejected characters can be flagged by the OCR system, and are therefore easily retraceable for manual correction.

3.5.3. Error rate.

The proportion of characters erroneously classified. Misclassified characters go by undetected by the system, and manual inspection of the recognized text is necessary to detect and correct these errors.

4. CONCLUSION

In this paper, we have discussed a survey of feature extraction and classification techniques for optical character recognition. A lot of research has been done in this field. Still the work is going on to improve the accuracy of feature extraction and classification techniques. However the different methods of feature extraction and classification discussed here are very effective and useful for new researchers. Template matching method which is easy to implement due to its algorithmic simplicity and higher degree of flexibility to the change of recognition target classes. Its recognition is strongest on monotype and uniform single column pages and it takes shorter time and does not require sample training but one template is only capable of recognizing characters of the same size and rotation. Neural network has ability to recognize characters through abstraction are great for scanned documents and damaged text.

REFERENCES

[1] R.O. Duda,P.T.Hart.D.G. Stork, Pattern Classification, 2nd ed.,A Wiley- Interscience Publication,2001.

[2] R.J.SCHALKOFF, Artificial Neural Networks, The McGraw-Hill Companies Inc., New York, 1997.

[3] Rice S., Nagy G., Nartker T., Optical Character Recognition:An Illustrated Guide to the Frontier, Springer, 1999.

[4] Mori S., Optical Character Recognition, Wiley- Interscience,1999.

[5] Parker J., Algorithms for Image Processing and Computer Vision, John Wiley & Sons Inc, 1996.

[6] Callan R., The Essence Of Neural Networks, Prentice-Hall,1999.

[7] Y. LeCun, B. Boaer, J.S. Denker, D. Henderson ,R.E.Howard, W. Hubbard and L.D. Jackel "Handwritten zip code recognition with multilayer networks" International Conference on Pattern recognition , pp. 35-44,1990.

[8] T.SITAMAHALAKSHMI, A.VINAY BABU, M.JAGADEESH,"Character Recognition Using Dempster-Shafer Theory combining Different Distance Measurement Methods," International Journal of Engineering Science and Technology, Vol. 2(5), 2010, 1177-1184.Mori S., Optical Character Recognition, Wiley- Interscience,1999.

[9] Kartar Singh Siddharth , Mahesh Jangid, Renu Dhir, Rajneesh Rani, "Handwritten Gurmukhi Character Recognition Using Statistical and Background Directional Distribution Features", *International Journal on Computer Science and Engineering (0975-3397)*, Vol. 3 No. 6 June 2011.

[10] Puneet Jhajj, D. Sharma, "Recognition of Isolated Handwritten Characters in Gurmukhi Script", *International Journal of Computer Applications* (0975-8887), Vol. 4, No. 8, 2010.

[11] M.-K. Hu, Visual pattern recognition by moment invariants,IRE Transactions Information Theory 8, 179–187,1962.

[12] Y. Li, Reforming the theory of invariant moments for pattern recognition, Pattern Recognition Letters 25, 723– 730, 1992.

[13] K. M. Mohiuddin and J. Mao, "A comparative study of different classifiers for handprinted character recognition," *Pattern Recognit. PracticeIV*, pp. 437–448, 1994

[14] "HMM–KNN word recognition engine for bank cheque processing," in *Proc. 14th Int. Conf. Pattern Recognit.*, Brisbane, Australia, 1998, pp. 1526–1529.

[15] B. Ripley, "Statistical aspects of neural networks," in *Networks on Chaos: Statistical and Probabilistic Aspects* , U. Bornndorff-Nielsen, J.Jensen, and W. Kendal, Eds. London, U.K.: Chapman & Hall, 1993.

[16] H. D. Block, B. W. Knight, and F. Rosenblatt, "Analysis of a four layer serious coupled perceptron," *Rev. Mod. Phys.*, vol. 34, pp. 135–152, 1962.

[17] B. Ripley, "Statistical aspects of neural networks," in *Networks on Chaos: Statistical and Probabilistic Aspects* , U. Bornndorff-Nielsen, J.Jensen, and W. Kendal, Eds. London, U.K.: Chapman & Hall, 1993.

[18] M. Bokser, "Omnifont technologies," *Proc. IEEE*, vol. 80, pp. 1066–1078, 1992.

[19]] Nafiz Arica and Fatos T. Yarman-Vural, "An Overview of Character Recognition Focused on Off-Line Handwriting**"** IEEE Transactions on systems, man, and cybernetics—Part C:APPLICATIONS AND REVIEWS, VOL. 31, NO. 2, MAY 2001.

[20] Amarjot Singh, Ketan Bacchuwar, and Akshay Bhasin, "A Survey of OCR Applications**"** *International Journal of Machine Learning and Computing, Vol. 2, No. 3, June 2012*