

A Review on Character Recognition Techniques

Richa Goswami
M.Tech Scholar
Poornima College of
Engineering, Jaipur

O.P. Sharma
Professor
Poornima College Of
Engineering, Jaipur

ABSTRACT

Character Recognition (CR) has been studied from the past several decades, and is still a demanding research topic in the field of pattern recognition and image processing. There is a large demand for Optical Character Recognition on hand written documents. Now days, the rapidly growing computational power enables the implementation of the present Character Recognition methodologies and creates a growing demand on many emerging application fields, which require further advanced methodologies. This paper provides a comprehensive review of existing works in handwritten character recognition based on soft computing technique during the past decade.

General Terms

Pattern Recognition

Keywords

Handwritten Character recognition, feature extraction, segmentation, training and recognition.

1. INTRODUCTION

Machine simulation of human functions has been an extremely challenging research field since the advent of digital computers in the field of image processing and pattern recognition [1]. In this paper, all types of machine recognition of character in various application domains have been covered by character recognition (CR). The intensive research in the field of CR is not only because it helps in postal address, bank check amount reading but also because it automates processing of bulk amount of papers, transfers data into machines and web interface to paper document.

In this overview, CR is used as a submissive term, which covers all types of machine recognition of characters in a variety of application areas. The overview shows an update for the state-of-the-art in the CR field, providing an emphasis on the methodologies essential for the increasing needs in newly rising areas, such as enhancement of multimedia databases and systems, electronic libraries etc, which engross handwriting data entry. The main approaches for offline cursive word recognition can be divided into 2 parts one is segmentation-based and the other one is holistic approach [2]. The holistic one is used in recognition where global features, taken out from the intact word image are considered and can be used for limited size vocabulary. As the extent of the vocabulary increases, the complexity of algorithms also increases linearly due to the need for a larger search space and a more complex pattern representation. In addition, the rates of recognition decrease rapidly due to the decrease in interclass variances in the feature space. On the other hand the segmentation based strategies, utilize bottom-up approaches, initializing from character or stroke level and moving towards producing a meaningful text. With the support of segmentation stage, the trouble of recognition of simple isolated characters or strokes is reduced, which can be handled for unlimited size of vocabulary.

The study explores the direction of the CR research, analyses the limitations of methodologies for the systems. The methodologies for CR can be classified based upon two main criteria: 1) the data acquisition process (on-line or off-line) and 2) the text type (machine-printed or handwritten). It is not a matter in which class the problem exists, in general, there are six major stages in the CR problem as shown in fig 1:

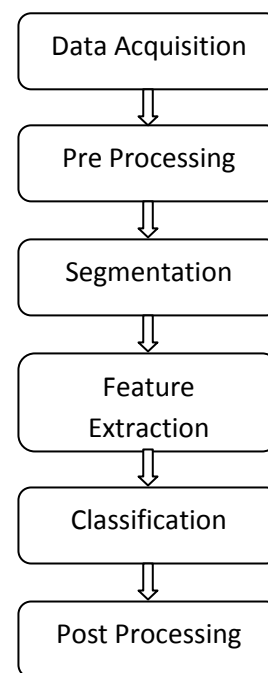


Fig.1 Flowchart for CR methodology

A number of techniques are available for feature extraction and training of CR systems each with its own superiorities and weaknesses. This paper is arranged as follows: section I describes the introduction, in section II evolution history of CR system is given, methodologies of CR have been discussed in Section III and the conclusion of the paper and the future scope of the technique is presented in section IV.

2. HISTORICAL EVOLUTION OF CR

Writing, which has been the most natural mode of accumulating, storing, collecting and transmitting information through the centuries, now serves not only for communication among humans but also serves for communication of humans and machines. The very first effort in the direction of CR was made by Turing who attempted to develop an aid for the visually handicapped [2]. The first character recognizer appeared in around 1940s. The early works were concentrated either upon machine-printed text or upon a small set of well-

separated hand-written text or symbols. Machine-printed CR generally used template matching and for handwritten text, low-level image processing techniques were used on the binary image to extract feature vectors, which were then fed to statistical classifiers [3]. A good survey of the CR techniques used until 1980s can be found in [4]. The period from 1980 - 1990 witnessed a growth in CR system development due to rapid growth in information technology [5].

Structural approaches were initiated in many systems in addition to the statistical methods [6]. The syntactic and structural approaches require efficient extraction of primitives [7]. Chan et al. [8] discussed a structural approach for recognizing on-line handwriting. The recognition process starts with a sequence of points from the user and then these points are used to take out the structural primitives. Different types of line segments and curves are incorporated in these primitives. But there existed an upper limit in the recognition rate, because the CR research was focused basically on the shape recognition techniques without using any semantic information. Historical review of CR research and development during 1980-1990 can be found in [9] for both on-line and off-line strategies, respectively.

After 1990, image processing techniques and pattern recognition were combined using artificial intelligence. Along with powerful computers and more accurate electronic equipments such as scanners, cameras and electronic tablets, there came in efficient, modern use of methodologies such as neural networks (NNs), hidden Markov models (HMMs), fuzzy set reasoning, and natural language processing. The 1990's systems for the machine-printed off-line [10] and restricted vocabulary size, user-dependent on-line handwritten characters [11] were satisfactory only for restricted applications. Although research on recognizing isolated handwritten characters has been quite successful, recognizing off-line cursive handwriting has been found to be a challenging problem. There is a large corpus of research on the application of character recognition in different domains, but no system to date has achieved the goal of system acceptability.

3. METHODOLOGIES OF CR SYSTEM

In this section the available methodologies to develop the stages of the CR system are presented. The CR approach varies a great extent, depending upon the nature of CR system and the methodology used. The literature review in the field of CR indicates that these hierarchical tasks are grouped in the stages of the CR for preprocessing followed by segmentation, representation, then training and recognition, and postprocessing. In several methods, some of the stages are combined or omitted; in others a feedback mechanism is used to update the output of each stage. The typical flowchart for CR methodology is shown in Fig 1 and is discussed as follows:

3.1 Data acquisition

The automatic character recognition systems advancement is evolved in two categories according to the approach of data acquisition:

- On-line character recognition systems
- Off-line character recognition systems

The character recognition off-line, captures the statistics from document through optical scanners or cameras whereas the recognition systems on-line, make use of the digitizers which directly captures writing through the order of the strokes, speed, pen-up and pen-down information.

3.2 Pre-processing:

The raw data, collected from the data acquisition stage, is subjected to variety of preliminary processing steps to make it usable in the descriptive stages of character analysis. The main objective of the pre-processing is to organise the information so that the subsequent CR task become simpler. It essentially enhances the image rendering it suitable for segmentation. The various Techniques performed on the image during pre-processing stage are discussed below:

1) noise reduction: Disconnected line segments, filled loops, bumps and gaps in lines etc are some variety of noises which are introduced by optical scanning devices. Also the other major problems are the distortions which include local variations, corners rounding, dilation, and erosion. It is required to eliminate these imperfections prior to the CR. A large number of existing noise reduction techniques can be categorized in three major groups [12].

a) Filtering: This aims to remove noise and diminish spurious points, usually introduced by uneven writing surface and/or poor sampling rate of the data acquisition stage. A range of spatial and frequency domain filters can be designed for this purpose. The basic 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. Filters can be designed for smoothing, sharpening, thresholding, removing slightly textured or colored background, and contrast adjustment purposes [13].

b) Morphological Operations: The basic idea behind the morphological operations is to filter the document image replacing the convolution operation by the logical operations. Various morphological operations can be designed to connect the broken strokes, decompose the joined strokes, smooth the contours, reduce the wild points, thin the characters, and extract the boundaries [14]. Therefore, morphological operations can be used successfully to take away the noise from the document images, due to poor quality of ink and document, as well as inconsistent hand movement.

c) Noise Modeling: Noise could be removed by several calibration techniques if a model for it were accessible. However, in most of the applications modelling of the noise is not probable. There is very little work on modeling the noise introduced by optical distortion, such as speckle, skew, and blur. Nevertheless, it is possible to assess the quality of the documents and remove the noise to a certain degree, as suggested in [15].

2) Normalisation: The most important part of the pre-processing stage is normalization, which attempts to eradicate some of the variations in the images, which do not affect the identity of the word. Normalization of the handwritten image from a scanned image includes some steps, which generally begin with cleaning of image, skew correction, line detection, slant and slope removal and character size normalization [15].

3) Compression: Space domain techniques are required for compression. Two important techniques are thresholding and thinning [16]. Thresholding reduces the storage requirements and increases the speed of processing by converting the gray-scale or colour images to binary image by taking a threshold value. Thinning extracts the shape information of the characters.

3.3 Segmentation

The pre processing stage yields a “clean” manuscript in the sense that a satisfactory amount of shape information, low noise on a normalized image, and high compression is obtained. The next step is to segment the document into its subcomponents. Segmentation is a significant stage because the level one can achieve in separation of lines, words or characters affects the recognition rate of the script directly. Segmentation is basically of two types: external segmentation, which is the separation of various writing units, such as words, sentences or paragraphs, and internal segmentation, which is the separation of letters, mainly in cursive written words.

3.4 Representation or Feature extraction

Image representation plays one of the most essential roles in a recognition system. The simplest way is to feed gray-level or binary images to a recognizer. However, in order to keep away the extra complexity and to raise the accuracy of the algorithms, a more compressed and characteristic representation is essential in most of the recognition systems. For this reason, a set of features is extracted for each class that helps to differentiate it from other classes while remaining invariant to characteristic differences inside the class [17]. In the following, hundreds of document image representation methods are categorized into three major groups.

1) *Global Transformation and Series Expansion*: A continuous signal generally contains more information than needs to be represented for the purpose of classification. This may be true for discrete approximations of continuous signals as well. One approach to represent a signal is by a linear combination of a sequence of simpler well-defined functions. The linear combination coefficients provides a compact encoding known as transformation or/and series expansion. Common transform and series expansion methods used in the CR field includes the following.

a) *Fourier Transforms*: The common procedure is to choose magnitude spectrum of the measurement vector as the features in an n-dimensional Euclidean space. One of the most attractive properties of the Fourier transform is the capability to identify the position-shifted characters, as soon as it observes the magnitude spectrum and ignores the phase. Fourier transforms have been useful to CR in many ways [18].

b) *Gabor Transform*: This is a distinction of the windowed Fourier transform. In this case, the window used is defined by a Gaussian function rather than a window of discrete size [19].

c) *Wavelets*: Wavelet transformation is a series expansion technique that allows one to represent the signal at different levels of resolution. The segments of document image, which may correspond to letters or words, are represented by wavelet coefficients, corresponding to a variety of levels of resolution. These coefficients are then fed to a classifier for recognition [20]

d) *Moments*: Moments, such as central moments, Legendre moments, and Zernike moments, form a compact representation of the original document image that make the process of recognizing an object scale, translation, and rotation invariant [21]. Moments are considered as series expansion representation since the original image can be completely reconstructed from the moment coefficients.

e) *Karhunen–Loeve Expansion*: This is an eigenvector analysis which attempts to reduce the dimension of the feature

set by creating new features that are linear combinations of the original ones. It is the only optimal transform in terms of information compression.

2) *Statistical Representation*: Representation of a document image by statistical distribution of points takes care of style variations to some extent. Although this type of representation does not allow the reconstruction of the original image, it is used for reducing the dimension of the feature set providing high speed and low complexity. The following are the major statistical features used for character representation:

a) *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.

b) *Crossings and Distances*: A popular statistical feature is the number of crossing of a contour by a line segment in a specified direction. In [22], the character frame is partitioned into a set of regions in various directions and then the black runs in each region are coded by the powers of two. Another study [23] encodes the location and number of transitions from background to foreground pixels alongside vertical lines throughout the word. Now, the line segments distance such as the upper and lower segment of the frame, from a given boundary, can also be used as statistical features. These features imply that a horizontal threshold is established above, beneath and through the centre of the normalized script. The value of that feature is decided by the number of times the script crosses the threshold. The obvious intention is to take the ascending and descending portions of the script.

c) *Projections*: Characters can be represented by projecting the pixel gray values onto lines in various directions. This representation creates a 1-D signal from a two-dimensional (2-D) image, which can be used to represent the character image [24].

3) *Geometrical and Topological Representation*: Various global and local properties of characters can be represented by geometrical and topological features with high tolerance to distortions and style variations. The representation of this type may also encode some knowledge about the structure of the object or may provide some knowledge as to what sort of components make up that object.

3.5 Classification

The classification stage is the decision making part of the recognition system. The performance of a classifier relies on the quality of the features. There are many existing Classical and soft computing techniques for handwriting classification. They are specified as:

(a) Classical Techniques:

Template matching

Statistical techniques

Structural techniques

(b) Soft Computing Techniques:

Neural networks (NNs)

Fuzzy- logic technique

Evolutionary computing techniques

1) *Template matching*: The simplest way of character recognition is based on matching the stored prototypes against the character or word to be recognized. Generally speaking,

matching operation determines the degree of similarity between two vectors (group of pixels, curvature, shapes etc.) in the feature space. Matching techniques can be studied in three classes Direct Matching, Deformable Templates and Elastic Matching and Relaxation Matching.

2) **Statistical Techniques:** Statistical decision theory is concerned with statistical decision functions and a set of optimality criteria, which maximizes the likelihood of the observed pattern given the model of a certain class. Statistical techniques are, mainly, based on three major assumptions:

i. Feature set distribution is Gaussian or in the worst case uniform,

ii. There are sufficient statistics available for each class,

iii. Given a group of images $\{I\}$, one is able to extract a set of features $\{f_i\} \in F$, $i \in \{1, \dots, n\}$, which represents each distinct class of patterns. The measurements taken from n -features of every word unit can be thought to represent an n -dimensional vector space and the vector, whose coordinates correspond to the taken measurements, shows the original word unit.

3) **Structural Techniques:** The recursive description of a complex pattern in terms of simpler patterns based on the shape of the object was the initial idea behind the creation of structural pattern recognition. These patterns are used to describe and classify the characters in the CR systems.

4) **Neural Networks (NN):** A neural network is defined as a computing architecture that consists of 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. It can easily adapt to changes in the data and learn the input signals characteristics, because of its adaptive nature. A neural network consists of 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 neural network architectures are equivalent to statistical pattern recognition methods [25]. Several approaches exist for training of neural networks [26]. These include the boltzman, error correction, hebbian and competitive learning. They cover binary and continuous valued input, in addition with supervised and unsupervised learning. On the other hand, neural network architectures can be classified into two major groups, namely, feed-forward and feedback (recurrent) networks. The most familiar neural networks used in the CR systems are the multilayer perceptron of the feed forward networks and the kohonen's Self Organizing Map (SOM) of the feedback networks.

3.6 Post processing

Post-processing stage is the last stage of the proposed recognition system [27]. It prints the corresponding recognized characters in the structured text form.

4. CONCLUSION

In this study, main approaches used in the CR field during the last decade are overviewed. Different Pre-processing, segmentation techniques and various classifiers with different features are also discussed. It is found that neither the structural nor the statistical information can represent a complex pattern alone. Therefore, one needs to combine statistical and structural information supported by the semantic information. NNs or HMMs are very successful in combining statistical and structural information for many

pattern recognition problems. They are comparatively resistant to distortions, but they have a discrete nature in the matching process, which may cause drastic mismatching. Template matching methods deal with a character as a whole in the sense that an input plane is matched against a template constrained on an X - Y plane. This makes the procedure very simple, and the complexity of character shape is irrelevant, but, it suffers from the sensitivity to noise and is not adaptive to variations in writing style. The characters are natural entities. They require complex mathematical descriptions to obey the mathematical constraint set of the formal language theory. Imposing a strict mathematical rule on the pattern structure is not particularly practical in CR, where intraclass variations are very large. In CR some features are tolerant to distortion and care of style variations, rotation, and translation should be taken to some extent. Design of the training set should be handled systematically, rather than putting up the available data in the set. Training sets should have considerable size and contain random samples, including poorly written ones.

5. REFERENCES

- [1] 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, pp.216-233, May 2001.
- [2] J. Mantas, "An overview of character recognition methodologies", Pattern Recognition, vol. 19, no. 6, pp. 425 - 430, 1986.
- [3] T. S. El-Sheikh and R. M. Guindi, "Computer recognition of arabic cursive script", Pattern Recognition, vol. 21, no. 4, pp. 293 - 302, 1988.
- [4] C. Tappert, C. Suen, and T. Wakahara, "The state of the art in online handwriting recognition", Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 12, pp. 787 -808, Aug. 1990.
- [5] Q. Tian, P. Zhang, T. Alexander, and Y. Kim, "Survey: omnifont-printed character recognition," in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series (K.-H. Tzou & T. Koga, ed.), vol. 1606 of Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, pp. 260-268, Nov. 1991.
- [6] Y. Ding, F. Kimura, Y. Miyake, and M. Shridhar, "Evaluation and improvement of slant estimation for handwritten words", in Document Analysis and Recognition, 1999. ICDAR '99. Proceedings of the Fifth International Conference on, pp. 753 -756, Sept. 1999.
- [7] S. M. Lucas, E. Vidal, A. Amiri, S. Hanlon, and J.-C. Amengual, "A comparison of syntactic and statistical techniques for off-line ocr", in Proceedings of the Second International Colloquium on Grammatical Inference and Applications, (London, UK), pp. 168-179, Springer-Verlag, 1994.
- [8] K.-F. Chan and D.-Y. Yeung, "Recognizing on-line handwritten alphanumeric characters through exible structural matching", Department of Computer Science, Volume 32, Issue 7 , Pages 1099–1114 ,1999.

- [9] S. Mori, C. Suen, and K. Yamamoto, "Historical review of ocr research and development", Proceedings of the IEEE, vol. 80, pp. 1029 - 1058, July 1992.
- [10] I. Bazzi, R. Schwartz, and J. Makhoul, "An omnifont open-vocabulary ocr system for english and arabic", Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 21, pp. 495 -504, June 1999.
- [11] J. Hu, S. G. Lim, and M. K. Brown, "Writer independent on-line handwriting recognition using an hmm approach", Pattern Recognition, vol. 33, no. 1, pp. 133 - 147, 2000.
- [12] Alessandro Vinciarelli, —"Noisy Text Categorizationl,", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.27, no.12, pp. 1882-1889,December 2005.
- [13] Ujjwal Bhattacharya, and B.B. Chaudhuri, "Handwritten Numeral Databases of Indian Scripts and Multistage Recognition of Mixed Numerals", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.31, no.03, pp.444-457, march 2009.
- [14] J. Yang and X. B. Li, "Boundary detection using mathematical morphology", Pattern Recognit. Lett., vol. 16, no. 12, pp. 1287–1296, 1995.
- [15] Salvador Espan˜a-Boquera, Maria Jose Castro-Bleda, Jorge Gorbe-Moya, and Francisco Zamora-Martinez "Improving Offline Handwritten Text Recognition with Hybrid HMM/ANN Models", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.33, no. 4, pp. 767-779, April 2011.
- [16] G.Pirlo and D. Impedovo, "Fuzzy-Zoning-Based Classification for Handwritten Characters", IEEE Transaction on pattern Recognition and Machine Intelligence, vol.19, no. 04, pp.780-785, August 2011.
- [17] I. S. Oh, J. S. Lee, and C. Y. Suen, "Analysis of class separation and combination of class-dependent features for handwriting recognition", IEEE Trans. Pattern Anal. Machine Intell., vol. 21, pp. 1089–1094, Oct. 1999.
- [18] S. S.Wang, P. C. Chen, andW. G. Lin, "Invariant pattern recognition by moment Fourier descriptor", Pattern Recognition., vol. 27, pp. 1735–1742, 1994.
- [19] Yuefeng Chen, Chunlin Liang, Lingxi Peng and Xiuyu Zhong, "A Handwritten Character Recognition Algorithm based on Artificial Immune", International Conference on Computer Application and System Modeling (ICCASM), pp. v12-273-276, 2010.
- [20] Yu Yang, "Handwritten Nepali Character Recognition Based on Wavelet Packet Transform and Artificial Immune System", International Conference on Computer Science and Service System, pp.442-445, August 2011.
- [21] Y. C. Chim, A. A. Kassim, and Y. Ibrahim, "Character recognition using statistical moments", Image Vis. Comput., vol. 17, pp. 299–307, 1999.
- [22] Muhammad Sarfraz1 and Ali Taleb Ali Al-Awami, "Arabic Character Recognition using Particle Swarm Optimization with Selected and Weighted Moment Invariants", Sixth International Conference on Signal Processing and its Application,ISSPA, pp. 1-4, May 2007.
- [23] C. Suresh Kumar ,Dr. T. Ravichandran , "Handwritten Tamil Character Recognition and Conversion using Neural Network ",International Journal on Computer Science and Engineering 2(7), 2261-2267, 2010.
- [24] Y. Tao and Y. Y. Tang, "The feature extraction of Chinese character based on contour information", in Proc. 5th Int. Conf. Document Anal. Recognit., Bangalore, India, pp. 637–640, 1999.
- [25] Ashoka H.N., Manjaiah D.H. ,Rabindranath Bera , "Feature Extraction Technique for Neural Network Based Pattern Recognition",International Journal of Computer Science & Engineering (IJCSE) 4(3), pp. 331-339, 2012.
- [26] Saurabh Shrivastava and Manu Pratap Singh, "Performance evaluation of feed-forward neural network with soft computing techniques for hand written English alphabets", Applied Soft Computing, vol. 11,pp. 1156–1182, 2011.
- [27] Shabana Mehfuz, Gauri Katiyar, "Intelligent Systems for Off-Line Handwritten Character Recognition: A Review" International Journal of Emerging Technology and Advanced Engineering ISSN 2250-2459, Volume 2, Issue 4, April 2012.