*International Journal of scientific research and management (IJSRM)*
*||Volume||4||Issue||07||Pages||4421-4428||2016||*
*Website: www.ijsrm.in ISSN (e): 2321-3418*

# Optical Character Recognition

***Pooja[1], Aakash[2]***
[1,2]Department of Computer Science Gateway Institute of Engineering & Technology (GIET), Deenbandhu Chhotu Ram University of Science & Technology (DCRUST), Sonepat
[1]ppoojapandita2@gmail.com
[2]aakash@gateway.edu.in

Abstract— Character Recognition (CR) has been an active area of research and due to its diverse applicable environment; it continues to be a challenging research topic. In this paper, we focus specially on off-line recognition of handwritten English words. The main approaches for off-line cursive word recognition can be divided into segmentation-based and holistic one. The holistic approach is used in recognition of limited size vocabulary where global features, extracted from the entire word image are considered. As the size 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. Additionally, the recognition rates decrease rapidly due to the decrease in interclass variances in the feature space. The segmentation based strategies, on the other hand, employ bottom-up approaches, starting from stroke or character level and going towards producing a meaningful text. With the cooperation of segmentation stage, the problem is reduced to the recognition of simple isolated characters or strokes, which can be handled for unlimited vocabulary

**Keywords—** Character Recognition, Feature extraction, PCA, ICA

## Introduction

Character recognition is one of the most successful applications of neural network technology. Handwriting recognition can be defined as the task of transforming text represented in the spatial form of graphical marks into its symbolic representation. In character recognition, printed documents are transformed into ASCII files for the purpose of editing, compact storage, fast retrieval through the use of computer. The recognition of character in a document becomes difficult due to noise, distortion, various character fonts and size, writing styles as handwriting of different persons is different. Many type of techniques for character recognition of several languages such as English, Hindi, Arabic, Chinese have been published but still recognition of characters using neural network is an open problem in terms of high recognition accuracy and minimum time of handwritten characters .

Earlier attempts for handwritten character recognition involved the extraction of representative features from the training data. These features were designed manually and special templates were created to detect them. Later on, efforts were made to automatically generate these features. This automatic generation of features improved the recognition rate and made design of a neural network for character recognition easier. Today, many researchers have been done to recognize characters, but the problem of interchanging data between human beings and computing machines is a challenging task. Even today, many algorithms have been proposed by many researchers so-that these characters can be easily recognize. But the efficiency of these algorithms is not satisfactory.

DOI: 10.18535/ijsrm/v4i7.10

Handwritten character recognition can be differentiated into two categories i.e. Online Handwritten character recognition and Offline Handwritten character recognition. On-line handwritten character recognition deals with automatic conversion of characters, which are written on a special digitizer, tablet PC or PDA where a sensor picks up the pen-tip movements as well as pen-up/pen-down switching. Off-line handwritten character recognition deals with a data set, which is obtained from a scanned handwritten document.

**Applications**

The most common application with example for the handwritten character recognition is as follows:

1. Documents processing: People wish to scan in a document and have the text of that document available in a word processor. For example HCR can be used for form processing. Forms are normally used for collecting the public information. Replies of public information can be handwritten in the space provided e.g. Tax Form, automatic accounting procedures used in processing utility bills.

2. Reading license plate numbers: Character recognition system is used widely for recognizing license plate numbers by traffic police e.g. Chinese traffic police uses OCR for reading license plate numbers while vehicle is in moving state.

3. Reading zip-codes for Post Office: Handwritten recognition system can be used for reading the handwritten postal address on letters. Offline handwritten recognition system used for recognition handwritten digits of postcode. OCR can be used to read this code and can sort mail automatically e.g. United State Postal Services (USPS) uses Multi line Optical Character Reader (MLOCR) locates the address block on a mail piece , reads the whole address, identifies the ZIP+4 code generates 9-digit bar code and sorts the mail to the correct stacker. The character classifier recognizes up to 400 fonts and the system can process up to 45,000 mail pieces per hour.

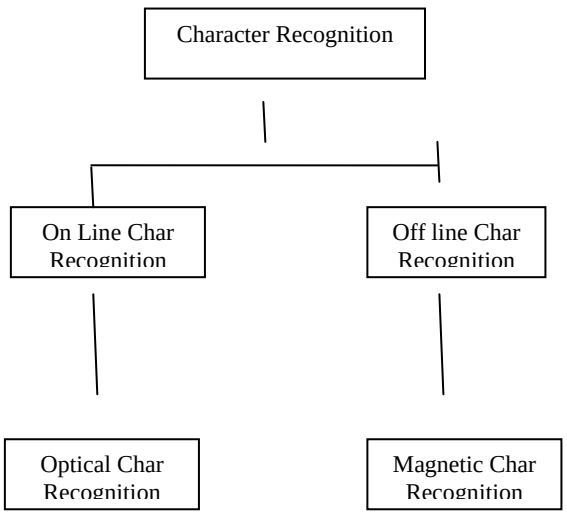
4. Reading Cheques in Banks: The OCR system is used for cheque reading in banks. Cheque reading is the very important commercial application of offline handwritten recognition. Handwritten recognition system plays very important role in banks for signature verification and for recognition of amount filled by user.

5. Signature Verification: OCR can be also used for signature verification to identify the person. Signature identification is the specific field of handwritten identification in which the writer is verified by some specific handwritten text. Handwritten recognition system can be used for identify the person by handwriting, because handwriting may be vary from person to person.

**Types of Character Recognition**

Character recognition can be classified into following two categories (Figure 1 below):

Online Character Recognition
Offline Character Recognition



DOI: 10.18535/ijsrm/v4i7.10

**Figure 1**: Character Recognition Classification

**1. On-line Character Recognition System**

On-line character recognition refers to the process of recognizing handwriting recorded with a digitizer as a time sequence of pen coordinates. In case of online handwritten character recognition, the handwriting is captured and stored in digital form via different means. Usually, a special pen is used in conjunction with an electronic surface. As the pen moves across the surface, the two- dimensional coordinates of successive points are represented as a function of time and are stored in order [18]. It is generally accepted that the on-line method of recognizing handwritten text has achieved better results than its off-line counterpart. This may be attributed to the fact that more information may be captured in the on-line case such as the direction, speed and the order of strokes of the handwriting.

The on-line handwriting recognition problem has a number of distinguishing features which must be exploited to get more accurate results than the online recognition problem

- It is adaptive: The immediate feedback is given by the writer whose corrections can be used to further train the recognizer.

- It is a real time process: It captures the temporal or dynamic information of the writing. This information consists of the number of pen strokes, the order of pen-strokes. The direction of the writing for each pen stroke and the speed of the writing within each pen stroke.

- Very little preprocessing is required. The operations such as smoothing, and feature extraction operations such as the detection of line orientations corners loops are easier and faster with the pen trajectory data than on pixel images.

- Segmentation is easy: Segmentation operations are facilitated by using the pen lift information particularly for hand printed characters.

- Ambiguity is minimal: The discrimination between optically ambiguous characters may be facilitated with the pen trajectory information

- On the other hand, the disadvantages of the on-line character recognition are as follows:

- The writer requires special equipment which is not as comfortable and natural to use as pen and paper.

- It cannot be applied to documents printed or written on papers punching is much faster and easier than handwriting for small size alphabet such as English or Arabic.

**2. Off-line Character Recognition System**

Off-line handwriting recognition refers to the process of recognizing words that have been scanned from a surface (such as a sheet of paper) and are stored digitally in grey scale format. After being stored, it is conventional to perform further processing to allow superior recognition.

The offline character recognition can be further grouped into two types:

- Magnetic Character Recognition (MCR)
- Optical Character Recognition (OCR)

In MCR, the characters are printed with magnetic ink. The reading device can recognize the characters according to the unique magnetic field of each character. MCR is mostly used in banks for check authentication. OCR deals with the recognition of

DOI: 10.18535/ijsrm/v4i7.10

characters acquiring by optical means, typically a scanner or a camera. The characters are in the form of pixelized images, and can be either printed or handwritten, of any size, shape, or orientation.

The OCR can be subdivided into handwritten character recognition and printed character recognition. Handwritten Character Recognition is more difficult to implement than printed character recognition due to diverse human handwriting styles and customs. In printed character recognition, the images to be processed are in the forms of standard fonts like Times New Roman, Arial, Courier, etc.

The drawbacks of the off-line recognizers, compared to on-line recognizers are summarized as follows:

- Off-line conversion usually requires costly and imperfect pre-processing techniques prior to feature extraction and recognition stages.
- They do not carry temporal or dynamic information such as the number and order of pen-on and pen-off movements, the direction and speed of writing and in some cases, the pressure applied while writing a character.
- They are not real-time recognizers.

Table 1 shows the comparison between online and offline handwritten character recognition.

**Table 1:** Comparison between online and offline handwritten characters

| Sr. No. | Comparisons | Online Char | Offline Char |
|---|---|---|---|
| 1 | Availability of Pen strokes | Yes | No |
| 2 | Raw data requirements | # samples/ seconds | # dots/ inch |
| 3 | Ways of writing | Digital Pen | Paper Document |
| 4 | Recognition Rate | Higher | Lower |
| 5 | Accuracy | Higher | Lower |

Phases of Handwritten Character Recognition

The process of handwritten character recognition can be divided into phases as shown in the figure 2 below

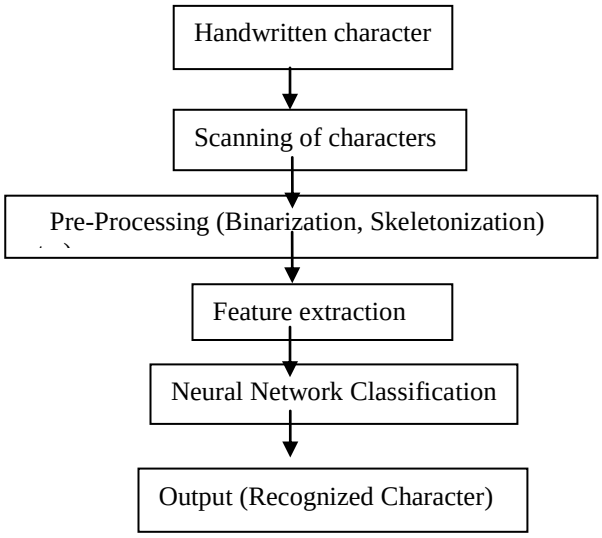



**Figure 2:** Block diagram for handwritten character recognition

**A. Pre-Processing**

Pre-processing is the name given to a family of procedures for smoothing, enhancing, Filtering, cleaning-up and otherwise massaging a digital image so that subsequent algorithm along the road to final classification can be made simple and more accurate.

Various Pre-processing Methods are explained below:

**A1. Binarization**

Document image Binarization (thresholding) refers to the conversion of a gray-scale image into a binary image. Two categories of thresholding:

- Global, picks one threshold value for the entire document image which is often based on an estimation of the background level from the intensity histogram of the image.
- Adaptive (local), uses different values for each pixel according to the local area information

**A2. Noise Removal**

The major objective of noise removal is to remove any unwanted bit-patterns, which do not have any significance in the output. Various filtering operations can be applied to remove noise e.g. Median Filter, Weiner filter etc.

**A3. Skeletonization**

Skeletonization is also called thinning. Skeletonization refers to the process of reducing the width of a line like object from many pixels wide to just single pixel. This process can remove irregularities in letters and in turn, makes the recognition algorithm simpler because they only have to operate on a character stroke, which is only one pixel wide. It also reduces the memory space required for storing the information about the input characters and no doubt, this process reduces the processing time too.

**A4. Smoothing**

The objective of smoothing is to smooth shape of broken and/or noisy input characters.

**A5. Thresholding**

In order to reduce storage requirements and to increase processing speed, it is often desirable to represent grey scale or color images as binary images by picking some threshold value for everything above that value is set to 1 and everything below is set to 0. Two categories of thresholding exist: Global and Adaptive. Global thresholding picks one threshold value for the entire document image. Adaptive thresholding is a method used for images in which different regions of the image may require different threshold values.

**A6. Normalization**

Normalization is a linear process. If the intensity range of the image is 50 to 180 and the desired range is 0 to 255 the process entails subtracting 50 from each of pixel intensity, making the range 0 to 130. Then each pixel intensity is multiplied by 255/130, making the range 0 to 255. Auto-normalization in image processing software typically normalizes to the full dynamic range of the number system specified in the image file format. The normalization process will produce iris regions, which have the same constant dimensions, so that two photographs of the same iris under different conditions will have characteristic features at the same spatial location [13]. Normalization methods aim to remove all types of variations during the writing and obtain standardized data [20]. For example Size normalization is used to adjust the character size to a certain standard. Methods of character recognition may apply both horizontal and vertical size normalizations.

**B. Feature Extraction**

Each character has some features, which play an important role in pattern recognition. Feature extraction describes the relevant shape information contained in a pattern so that the task of classifying the pattern is made easy by a formal procedure. In pattern recognition and in image processing, feature extraction is a special form of dimensionality reduction. The main goal of feature extraction is to obtain the most relevant information from the original data and represent that information in a lower dimensionality space. When the input data to an algorithm is too large to be processed and it is suspected to be redundant (much data, but not much information) then the input data will be transformed into a reduced representation set of features (also named features vector). Transforming the input data into the set of features is called feature extraction. If the features extracted are carefully chosen it is

expected that the features set will extract the relevant information from the input data in order to perform the desired task using this reduced representation instead of the full size input.

**Feature extraction techniques**

The techniques that are used for feature extraction are:

**1. Principal component analysis (PCA)**

PCA is a mathematical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of uncorrelated variables called principal components. The number of principal components is less than or equal to the number of original variables. This transformation is defined in such a way that the first principal component has as high a variance as possible (that is, accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it be orthogonal to (uncorrelated with) the preceding components. Principal components are guaranteed to be independent only if the data set is jointly normally distributed. PCA is sensitive to the relative scaling of the original variables. Depending on the field of application, it is also named the discrete Karhunen–Loève transform (KLT), the Hotelling transform or proper orthogonal decomposition (POD).

**2. Independent Component Analysis (ICA)**

ICA is a statistical technique that represents a multidimensional random vector as a linear combination of nongaussian random variables ('independent components') that are as independent as possible. ICA has many applications in data analysis, source separation, and feature extraction.

**3. Corner detection**

Corner detection is an approach used within computer vision systems to extract certain kinds of features and infer the contents of an image. Corner detection is frequently used in motion detection, image matching, video tracking, 3D modeling and recognition. A corner can be defined as the intersection of two edges. A corner can also be defined as points for which there are two dominant and different edge directions in a local neighborhood of the point. A simple approach to corner detection in images is using correlation, but this gets very computationally expensive and suboptimal.

**4. Fourier Descriptor**

Fourier transformation is widely used for shape analysis [21, 22]. The Fourier transformed coefficients form the Fourier descriptors of the shape. These descriptors represent the shape in a frequency domain. The lower frequency descriptors contain information about the general features of the shape, and the higher frequency descriptors contain information about finer details of the shape. Although the number of coefficients generated from the transform is usually large, a subset of the coefficients is enough to capture the overall features of the shape. The high frequency information that describes the small details of the shape is not so helpful in shape discrimination, and therefore, they can be ignored. As the result, the dimensions of the Fourier descriptors used for capturing shapes are significantly reduced.

**Representation of Character Features**

After extracting the features, the data should be represented in one of two ways, either as a boundary or as a complete region. When the focus is on external shape characteristics such as corners and variation then boundary representation is appropriate. While regional representation is appropriate when the focus is on internal properties such as textures or skeletal shape. In some applications like character recognition these representations coexist, which often require algorithm based on boundary shape as well as skeletons and other internal properties. In terms of character recognition descriptors such as holes and bays are powerful features that help differentiate one part of the character from another.

This description also called feature selection, deals with extracting features which results in some quantitative information of interest or features that are basic for differentiating one class of objects from another.

## V. Conclusion

Recognition approaches heavily depend on the nature of the data to be recognized. The recognition process needs to be much efficient and accurate to recognize the characters written by different users. As neural network is used here for recognition of offline English character images and it has been seen that recognition increases, although at a slow rate. Also some characters like I & J are similar, so the recognition system gives sometimes bad results for similar character. Also it is based on the handwriting style e.g. G may be written as  or  . This may also create problem sometimes. Also it can be concluded that system is not stable. Every time it gives different results. This may be due to the number of character set used for training was reasonably low. As the network is trained with more number of sets, the accuracy of recognition of characters will increase definitely. It can be concluded that the work successfully does the character recognition. It has the limitation that it performs the training as well as testing at a slow rate. But from the above results, it may be said that accuracy may be achieved better if number of training set is taken larger and also if better image processing techniques are considered.

## References


1. Anita Pal, Dayashankar Singh "Handwritten English Character Recognition Using Neural Network", International Journal of Computer Science & Communication Vol. 1, No. 2, July-December 2010, pp. 141-144
2. C. Suresh kumar, Dr. T. Ravichandran, "Handwritten Tamil Character Recognition Using RCS Algorithm", International Journal of Computer Applications (0975 – 8887) Volume 8– No.8, October 2010
3. Rashad Al-Jawfi, "Handwriting Arabic Character Recognition LeNet Using Neural Network", The International Arab Journal of Information Technology, Vol. 6, No. 3, July 2009
4. Srinivasa Kumar DeviReddy, Settipalli Appa Rai, "Hand written character recognition using back propagation network", Journal of Theoretical and Applied Information Technology,2005-2009.
5. Cheng-Lin Liu, "Normalization-Cooperated Gradient Feature Extraction for Handwritten Character Recognition", IEEE Transaction on pattern analysis and machine intelligence, vol. 29, no. 8, Aug. 2007
6. Hadar I. Avi-Itzhak, Thanh A. Diep, and Harry Garland, "High Accuracy Optical Character Recognition Using Neural Networks with Centroid Dithering", in IEEE Transactions on pattern analysis and machine intelligence, vol. 17, no. 2, Feb.1995.
7. Youfu Wu, Yongwu Wu, Gang Zhou, Jing Wu, "Recognizing Characters Based on Gaussian-Hermite Moments and BP Neural Networks", in International Conference on Intelligent Computation Technology and Automation, 2010, ISBN 978-0-7695-4077-1/10.
8. Birijesh K. Verma – "Handwritten Hindi Character Recognition Using Multilayer Perceptron and: Radial Basis Function Neural Networks", in IEEE International conference on neural networks, Vol. 4, pp. 2111-2115, Nov.1995.
9. Janusz k Starzyk and Nasser Ansari – "Feedforward Neural Network for Handwritten Character Recognition", in IEEE symposium on circuit and systems, 1992.
10. Simon Haykin, Neural Networks, A Comprehensive Foundation, Pearson Education, Inc., 2004.

11. Rafael C. Gonzalez, Richard E. Woods, Digital Image Processing, 2nd edition, Pearson Education.
12. http://www.igi.tugraz.at/lehre/EW/tutorials/nn-ocr/nn-ocr.pdf
13. http://homepages.cae.wisc.edu/~ece539/project/f03/sarkar-rpt.pdf
14. http://en.wikipedia.org/wiki/Normalization/(image_processing)
15. http//en.wikipedia.org/wiki/Optical_character_recognition
16. Mark S. Nixon, Alberto S. Aguado, Feature Extraction and Image Processing, Newnes Publisher, 2002 ISBN 0 7506 5078 8.
17. http://www.mathworks.com/matlabcentral/fileexchange/18169-optical-character-recognition-ocr
18. R. Plamondon and S. N. Srihari, "On-line and off-line handwritten recognition: a comprehensive survey",IEEE Transactions on PAMI,Vol. 22(1), pp. 63–84, 2000.
19. Jean R. Ward and Thedore Kuklinski, "A Model for Variability Effects in Hand-writing Character Recognition Systems" in IEEE Trans. Sys. Man. Cybernetics. Vol.18 No 3 pp438-451, 1988.
20. W.Guerfaii and R. Plamondon, "Normalizing and Restoring On-line Handwriting", Pattern Recognition, Vol 26 No3pp 418-431, 1993.
21. G. G. Rajput, Rajeswari Horakeri, Sidramappa Chandrakant, "Printed and Handwritten Kannada Numeral Recognition Using Crack Codes and Fourier Descriptors Plate" in International Journal of Computer Application(IJCA) on Recent Trends in Image Processing and Pattern Recognition (RTIPPR), 2010 pp 53-58.
22. G. G. Rajput, S. M. Mali, "Marathi Handwritten Numeral Recognition using Fourier Descriptors and Normalized Chain Code" in International Journal of Computer Application(IJCA) on Recent Trends in Image Processing and Pattern Recognition (RTIPPR, 2010 pp 141-145