INTERNATIONAL JOURNAL OF
RESEARCH IN COMPUTER
APPLICATIONS AND ROBOTICS
ISSN 2320-7345

IDENTITY STRING GENERATION FOR OFFLINE HANDWRITTEN HINDI CHARACTER RECOGNITION

Minu Babu[1]

[1]*MTtech Scholar, Department of Computer Science and Engineering, Federal Institute of Science and Technology,*
Mahatma Gandhi University, Kottayam, Kerala
minubabu4@gmail.com

Abstract

Handwritten character recognition (HCR) has been one of the fascinating and challenging research areas in field of image processing and pattern recognition in the recent years. In general, handwritten character recognition classified into two types as off-line handwritten character recognition and on-line handwritten character recognition. In off-line HCR, the writing is usually captured optically by a scanner and complete writing is available as an image. But, in the on-line HCR the two dimensional coordinates of successive points of character is captured in real time and are represented as a function of time. A typical offline HCR system consists of several stages such as image acquisition, pre-processing, segmentation, feature extraction, classification, recognition and post processing. In Image acquisition, the recognition system considers a scanned image as an input. The pre-processing is a series of operation performed on the scanned input image. It essentially enhances the image quality by eliminating the unwanted regions and enhances the character. This will produce a character image suitable for segmentation. In the segmentation stage is employed in the case of complete word recognition. Here, an image of sequence of characters is decomposed into sub-images of individual character. In the proposed system, the pre-processed input image is uniformly resized into a fixed size of 90x60pixels. Then, it is subdivided into 54 equal divisions, each of size 10x10 pixels. Then, features are extracted from each division through a diagonal feature extraction scheme. Finally, 54 features are extracted for each character. Each of these features is then converted to 7-bit binary format. Feature which are extracted from pervious stage are used to form a 378-bit identity string, used to recognize the character image using neural network.

Keywords*:* Handwritten character recognition, Pre-processing, Feature extraction, Diagonal feature extraction, Classification, Neural network, Identity string.

1. Introduction

Machine simulation of human functions has been a very challenging research field since the advent of digital computers [1]. In some areas, which require certain amount of intelligence, such as number crunching or chess playing, tremendous improvements are achieved. On the other hand, humans still outperform even the most powerful computers in the relatively routine functions such as vision. Machine simulation of human reading is one of these areas, which has been the subject of intensive research for the last three decades, yet it is still far from the final frontier. Character Recognition (CR) [1] has been extensively studied in the last half century and progressed to a level, sufficient to produce technology driven applications. Now, the rapidly growing computational power enables the implementation of the present CR methodologies and also creates an increasing demand on many emerging application domains, which require more advanced methodologies. In general there are seven major stages in the CR problem as shown in Figure 1. Handwriting has changed tremendously over time and, so far, each technology-push has contributed to its expansion. The printing press and typewriter opened up the world to formatted documents, increasing the number of readers that, in turn, learned to write and to communicate.

Computer and communication technologies such as word processors, fax machines, and e-mail are having an impact on literacy and handwriting. Newer technologies such as personal digital assistants (PDAs) and digital cellular phones will also have an impact. All these inventions have led to the fine-tuning and reinterpreting of the role of handwriting and handwritten messages.

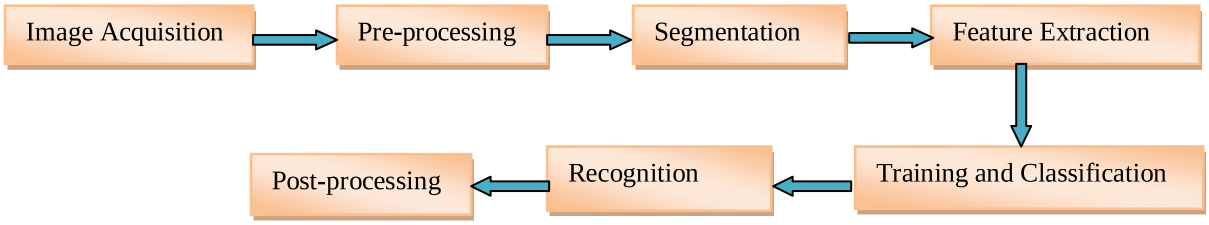

Figure 1: Different stages in Character Recognition

Widespread acceptance of digital computers seemingly challenges the future of handwriting. However, in numerous situations, a pen together with paper or a small notepad is much more convenient than a keyboard. Several types of analysis, recognition, and interpretation can be associated with handwriting. Some of the important applications of offline handwritten character recognition are Cheque Reading, Postcode Recognition, Form processing, Signature verification etc.

2. Literature Survey

Writing, this has been the most natural mode of collecting, storing and transmitting the information through the centuries. Now serves not only for the communication among humans, but also, serves for the communication of humans and machines. The intensive research effort on the field of CR was mainly due to its efficient applications such as the automatic processing of bulk amount of papers, transferring data into machines and web interface to paper documents.

The history of character recognition can be traced as early as 1900, when the Russian Scientist attempted to develop an aid for visually handicapped. The first character recognizers appeared in the middle of the 1940s with the development of the digital computers [2]. The early work on the automatic recognition of characters has been concentrated either upon machine printed text or upon small set of well-distinguished handwritten text or symbols. Machine-printed CR systems in this period generally used template matching in which an image is compared to a library of images. For handwritten text, low level image processing techniques have been used on the binary image to extract feature vectors, which are then fed to statistical classifiers. The studies until 1980 suffered from the lack of powerful computer hardware and data acquisition devices. With the explosion on the information technology, the previously developed methodologies found a very rapid growth in many application areas, as well as CR system development [3]. Structural approaches were initiated in many systems in addition to the statistical methods. These systems broke the character image into a set of pattern primitives such as lines and curves. Then rules were developed to determine which character most likely matched the extracted primitives. The real progress on CR systems is achieved during this period. The new development tools and methodologies were developed. Researchers developed complex CR algorithms, which receive high-resolution input data.

2.1 Existing Character Recognition Systems

The available CR systems are classified according to the data acquisition techniques and the text type. The progress in CR methodologies evolved in two categories according to the mode of data acquisition, as online and offline character recognition systems. The problem of recognizing handwriting recorded with a digitizer, as a time sequence of pen coordinates is known as on-line character recognition. The digitizers are mostly electromagnetic-electrostatic tablets, or pressure-sensitive tablets which send the coordinates of the pen tip to the host computer at regular intervals. The feature of online HCR is it is a real time process, it is adaptive in real time, it captures the temporal and dynamic information of the pen trajectory, very little pre-processing is required and segmentation is easy. The disadvantages of online HCR is that the writer requires special equipment, which is not as comfortable as pen and paper, it cannot be applied to documents printed or written on papers, punching is much faster and easier, the available systems are slow and recognition rates are low for handwriting that is not neat. The examples of online HCR is pen based computers, educational software for teaching handwriting, signature verifiers etc.

Off-line character recognition is known as Optical Character Recognition (OCR), because the image of writing is converted into bit pattern by an optically digitizing device such as optical scanner or camera. The recognition is done on this bit pattern data for machine-printed or hand-written text. The research and development is well progressed for the recognition of the machine-printed documents. In recent years, the focus of attention is shifted towards the recognition of hand-written script. The drawbacks of offline HCR is that it usually requires costly and imperfect pre-processing techniques prior to feature extraction and recognition stages, has lower recognition rates compared to on-line recognition. The major applications of offline HCR is large-scale data processing such as postal address reading, check sorting, office automation for text entry, automatic inspection and identification. Also, it is a very important tool for creation of the electronic libraries.

Considering the text type, CR systems are of two types as hand-written and machine-printed CR systems. Machine-printed text includes the materials such as books, newspapers, magazines, documents and various writing units in the video or still image. The available systems yield as well as 99% recognition accuracy [4]. However, the recognition rates of the commercially available products are very much dependent on the age of the documents, quality of paper and ink. Hand-written character recognition systems have still limited capabilities depends on the nature of writing.

3. Proposed Recognition System

A handwritten Hindi character recognition system using neural network is the proposed system. The different stages of the proposed system are shown in Figure 1.

3.1 Image Acquisition and Pre-processing

The image captured through a scanner is to be pre-processed first as in Figure 2. The main objectives of pre-processing are noise reduction, normalization of the data, and compression in the amount of information to be retained. Available noise reduction techniques can be categorized in three major groups [5] as filtering, morphological operations and noise modelling. Normalization methods aim to remove the variations of the writing and obtain standardized data. Compression for character recognition requires space domain techniques for preserving the shape information.

3.2 Feature Extraction

The system uses a diagonal feature extraction scheme for the recognizing off-line handwritten Hindi characters. In the feature extraction process (Figure 3), resized individual character of size 90x60 pixels is further divided into 54 equal zones, each of size 10x10 pixels. The features are extracted from the pixels of each zone by moving along their diagonals. This procedure is repeated for all the zones leading to extraction of 54 features for each character. These features are converted into identity bit string of size 378. Feature is extracted from each zone of size 10x10 is converted into 7 bit string so, there are 54 zone in each character image so, 54x7=378 bit identity string is used to represent each character image.

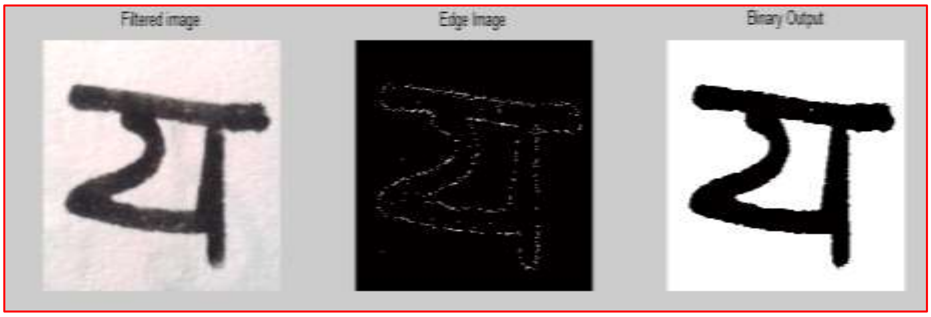

Figure 2: Pre-processing of inputted character image

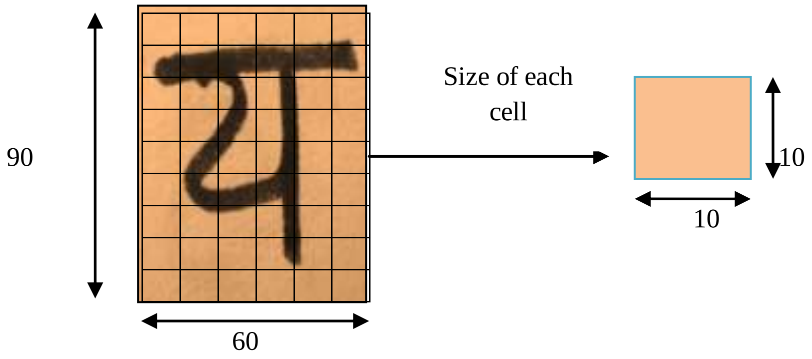

Figure 3: Feature Extraction of character image

3.3 Training, Classification and Recognition

Using these identity strings, the classifier is trained. An artificial neural network as the backend is used for performing classification and recognition tasks. In the off-line recognition system, the neural networks have emerged as the fast and reliable tools for classification towards achieving high recognition accuracy. 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. Because of its adaptive nature, it can adapt to changes in the data and learn the characteristics of input signal. The neural network architectures can be classified into two major groups: feed-forward and feedback (recurrent) networks. NN learning is of two types: supervised and unsupervised learning. The most common neural networks used in the CR systems are the multilayer perceptron of the feed forward and feedback networks. NN has a two level detection scheme: the first level is for detection of sub patterns and the second level is for detection of characters. Handwritten character recognition is not a simple task. In recognition step, using a fitness function the identity string difference between unknown character and which are store in data base are calculated. Character recognition is complex task, even after writing people are not able to understand but he/she written. So to reach 100% accuracy is very difficult job. Many researchers have done lots of work in this field but, 100% accuracy in not achieved.

3.4 Post processing

While pre-processing tries to “clean” the document in a certain sense, it may remove important information. The lack of context information during the segmentation stage may cause even more severe and irreversible errors, since it yields meaningless segmentation boundaries. The result obtained from the output of the recognition stage can be further processed through parsing in the post processing stage to increase the recognition rates. The character recognition result is shown in Figure 4.

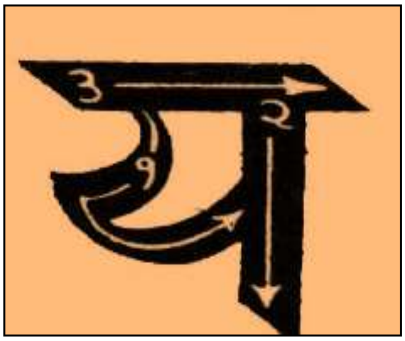

Figure 4: Character is recognized

4. Conclusion

A simple off-line handwritten Hindi script recognition system using a feature extraction method, namely, diagonal feature extraction is used in the above system. Identity bit string from the 54 feature which in extracted using the diagonal based feature extraction technique. It has been found that recognition of handwritten characters is very difficult task. Following are main reasons for difficulty in recognition of Hindi characters are:

- Some Hindi characters are similar in shape.
- Different or even the same writer can write differently depending upon pen or pencil.
- The character can be written at different location on paper or its window.
- Character can be written in different fonts.
- Difficult to create the training data set.

The success of any recognition system is depends on feature and classifier which is used to classify the unknown input to well define class. But as the concern of handwritten character each person write character in its own way, so well define structure is not applicable in this type of problem. So it needs more complex method like mutation, etc to make recognition of languages more productive.

References

1) Nafiz Arica and Fatos T. Yarman Vural, 1999, **“**An Overview of Character Recognition Focused On Off-line Handwriting”, Student Member, IEEE, Senior Member, IEEE IN.

2) L. D. Earnest and Amsterdam, 1963 “Machine Reading of Cursive Script”, in *Proc. IFIP Congress*, pp.462-466.

3) V. K. Govindan, A. P. Shivaprasad, 1990 “Character recognition- A review”, *Pattern Recognition* vol.23, no.7, pp.671-683.

4) S. S. Wang, P. C. Chen and W. G. Lin, 1994 “Invariant Pattern Recognition by Moment Fourier Descriptor”, *Pattern Recognition*, vol.27, pp.1735-1742.

5) M. Sonka, V.Hlavac and R. Boyle, 1999, “Image Processing, Analysis and Machine Vision*”*, 2nd Ed. *PWS Publishing, Brooks/Cole Pub. Company.*

Minu Babu – Currently doing her final year post graduation in Computer Science from Mahatma Gandhi University, Kottayam, Kerala. She took graduation from the same university in 2012 with distinction. This work is done as a part of her PG mini project. She is very much interested in Pattern Recognition, Image Processing, Machine Learning and Artificial Intelligence areas. Recently, she is interested in Emotion Recognition from speech and doing her main project in this area.