

# Handwritten Text Recognition

<sup>[1]</sup> M.Sivasankari & <sup>[2]</sup> Dr.S.P.Victor

<sup>[1]</sup> *Research Scholar (M.Phil Computer Science)*

<sup>[2]</sup> *Associate Professor & Head*

*Department of Computer Science, St.Xavier's College (Autonomous), Palayamkottai.*

**Abstract-**This text recognition lays emphasis on technique to recognize handwritten English characters using MATLAB. Here a novel solution is proposed for performing English character recognition. Handwritten character can be converted into the text. It may be either scanned document or paint image. Here the application is alphabetic letters are assigned with numerals accordingly. The new venture in this topic is to make off-line handwritten text recognition more accurate & challengeable. Here what is uploaded will come in the form of original text. Pursued by the preprocessing technique, feature extraction and recognition. The approach can be utilized in MATLAB for recognizing off-line handwritten English characters.

**Index Terms** — text recognition, handwritten, MATLAB

## INTRODUCTION

Recognition is regarded as a basic attribute of human beings, as well as other living organisms. A pattern is the description of an object. Performing the acts of recognition in every walk of our lives, such as recognition of objects around us, recognizing a friend in a crowd, recognizing voice of a known individual and recognizing handwritten characters.

According to the nature of the patterns to be recognized, we may divide our acts of recognition into two major types, namely, the recognition of concrete items and the recognition of abstract items. The concrete items of recognition include spatial patterns like characters, finger prints, weather maps, objects etc and temporal patterns like speech waveforms, electrocardiograms, target signatures, and time series Tou et al (1974). An example of recognition of an abstract item is the recognition of solution to a problem when your eyes and ears are closed. However we are concerned only with recognition of concrete items. Recognition of concrete patterns by human beings may be considered as a physiological problem which involves a relationship between a person and a physical stimulus. When a person perceives a pattern, he makes an inductive inference and associates this perception with some general concepts or clues which he has derived from his past experience. Human recognition in reality is a question of estimating the relative odds that the input data can be associated with one of a set of known statistical populations which depend on our past experience and which form the clues and the a priori information for recognition. Thus, the problem of pattern recognition may be regarded as one of discriminating the input data not between individual patterns but between populations, via the search for features or invariant attributes among members of a population.

Hence, Pattern Recognition can be defined as the classification of input data into identifiable classes via the extraction of significant features or attributes of the data from a background of irrelevant details. A Character recognition system is a pattern recognition system which receives scanned preprocessed data as input and identifies the name of the character by comparing it with the prototype characters already in the memory by Annadurai et al (2007). The study of pattern recognition problems may be logically divided into two major categories Ullman (1982). The study of the pattern recognition capability of human beings and other living organisms.

In this paper, the impart main emphasis on word recognition & hence assume that the handwritten texts are already segmented into words segment. Characters are limited to be digits on account of time constraints. Since it requires less image processing & data collection efforts. These handwritten texts are varied. Using the lexicon text is taken out the possibility towards correcting recognition.

## I. LITERATURE REVIEW

The Literature study is in category namely research work in character recognition.

Since 1980 a large amount of work has been carried out in character recognition. In the early 80's most of the works are related to either printed word or handwritten characters recognition only. In the mid of 80's Nagy et al (1987) demonstrated a heuristic algorithm for assigning alphabetic identity for symbols in a textual context, and a rule based contextual post processor for Devanagari recognition was suggested by Sinha (1985). Much of the work upto 1987 has been reported in Govindan's thesis (1988). The following are some of the recent works in this area. Kahan et al (1987) described the state of the system (OCR during 1987) that recognizes printed text of various fonts and sizes to any extent. Cheng et al (1985) discussed a method based on the concept of fuzzy set for handwritten Chinese character recognition. The strokes of the characters and its type and location features are concentrated to recognize the characters using fuzzy entropy. The state-of-the-art of Methodology and algorithms of fuzzy sets for real world use of Fuzzy sets in Pattern Recognition is described by Pedrycz (1990). The recognition and classification are provided with examples. Recognition of handwritten and machine printed text for postal address interpretation using various algorithms are discussed by Srihari (1992, 1993). Current state-of-the-art of new generation of handwriting recognition system with lessons learned and

difficulties involved and the challenges ahead are given in Suen et al (1980, 1993). The conventional techniques in Japan along with state-of-the-art and remaining problems are suggested in Wakahara (1993). New promising approaches are enumerated from the view point of general pattern recognition methodology. Sinha et al (1993) presented a hybrid contextual algorithm designed to read real-life documents printed in varying fonts of any size with string matching. Three passes are discussed in this paper. In the first pass character hypothesis is generated using dynamic contour warping classifier and the recognition is done based on transient statistical information collected. In the second pass word hypothesis is generated using hybrid contextual text processing and the recognition is done using Viterbi algorithm Hull et al (1982) and a string matching concept. In the third pass the word hypothesis is verified using cost criterion based on statistics and language heuristics.

**II. METHODOLOGY**

An overview of the methodologies adopted in this research work is presented below. A system is developed to recognize handwritten English characters using MATLAB, for a subset of the English characters.

**STEPS IN HANDWRITTEN CHARACTER RECOGNITION**

The major steps involved in handwriting recognition system suggested by Annadurai[7] are shown in Figure 2.

The steps involved are

1. Preprocessing
2. Feature Extraction
3. Recognition

**1. PREPROCESSING**

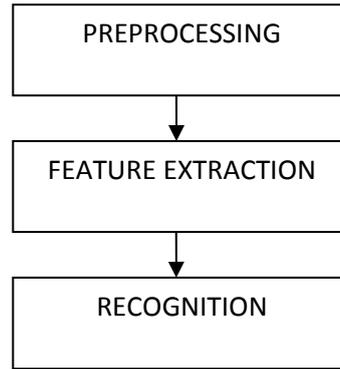
Preprocessing is concerned mainly with the reduction of these kinds of noise and variability in the input. The number and type of preprocessing algorithms employs on the scanned image depend on many factors such as paper quality, resolution of the scanned image, the amount of skew in the image and the layout of the text. Various preprocessing operations are performed prior to recognition to enhance the quality of the input image.

**2. FEATURE EXTRACTION**

Feature Extraction is the problem of extracting from the preprocessed data, the information, which is most relevant for classification purposes, in the sense of minimizing the within-class pattern variability, while enhancing the between-class pattern variability.

**3. RECOGNITION**

The final goal of character recognition is to obtain the class codes of character patterns. On segmenting character patterns or words from document images, the task of recognition becomes assigning each character pattern or word to a class out of a pre-defined class set. Pattern classification has been the main theme of pattern recognition field and is often taken as a synonym for "Pattern Recognition (PR)". Traditionally, pattern recognition techniques are classified as template and feature-based approach Pal et al(2004).



The framework of this method is shown in fig. 1

**III. EXPERIMENTAL RESULTS**

To assess the performance of each method. Here present some preliminary results under compression. Note that the practical choice of parameters can further be optimized in order to improve the results.

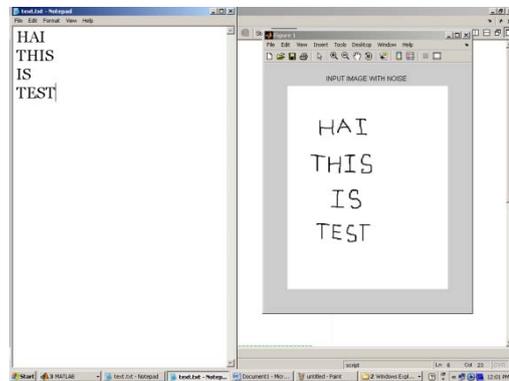


Figure 2. Shows the signature is realistic in practice

| MODEL                | RECOGNITION RATE |
|----------------------|------------------|
| OCTAL GRAPH APPROACH | 82%              |
| HMM                  | 92%              |
| SHMM                 | 97%              |
| PROPOSED METHOD      | 98%              |

Figure 2(a).Shows the comparison table

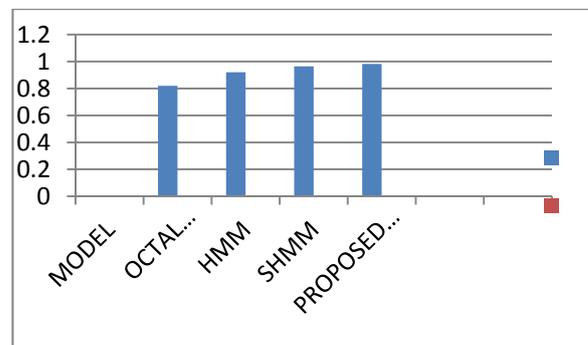


Figure 3. Experimental results show the proposed method is feasible in practice

#### IV. CONCLUSION

In this chapter the recognition of English Characters was improved using MATLAB to get the maximum possible efficiency. Significant increase in accuracy levels has been found with comparison of our method with the others for character recognition. The experimental results

It shows that the accuracy is really improved than the previous study. With the addition of sufficient pre processing the approach offers a simple and fast structure for fostering a full OCR system.

#### V. SCOPE FOR FUTURE WORK

The work reported in this thesis can be extended in the following directions. It is basically necessary to develop an OCR for multilingual script for a country where more than 10 languages are in use officially. For example if there is a document where two language scripts are available, then one need not separate two scripts into two different files and feed them into two OCR's. Using our approach one could develop an OCR for two languages, Tamil and English in the same document. These scripts could edit later with an editor too.

#### REFERENCES

- [1]. Abhuhaiba, I. S. I. and Ahmed, P. "A Fuzzy Graph theoretic approach to recognize the totally unconstrained handwritten Numerals", Pattern Recognition, Vol. 26, No. 9, pp. 1335-1350, 1993.
- [2]. Agui, T. and Nagahashi, N. "A description method of hand-printed Chinese Characters", IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 1, pp. 20-24, 1979.
- [3]. Alessandro Vinciarelli, Samy Bengio and Horst Bunke, "Offline Recognition of Unconstrained Handwritten Texts using HMMs and Statistical Language Models", IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 26, No. 6, pp. 709-720, 2004.
- [4]. Ali, F. and Pavlidis, T. "Syntactic recognition of handwritten numerals", IEEE Trans. Systems Man Cybernetics, Vol. 7, pp. 537-541, 1977.
- [5]. Almuallim, H. and Yamaguchi, S. "A method of recognition of Arabic cursive handwriting", IEEE. Trans. Pattern Analysis Machine Intelligence, Vol. 9, pp. 715-722, 1987.
- [6]. Amin, A., "Off-line Arabic Character Recognition: The State of The Art", Pattern Recognition, Vol. 19, No. 2, pp. 182-186, 1997.
- [7]. Annadurai, S. and Shanmugalakshmi, R. "Fundamentals of digital image processing", Pearson Education, 2007.
- [8]. Anoop Namboodiri, M. and Anil Jain, K. "Online Handwritten Script Recognition", IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 26, No. 1, pp. 124-130, 2004.
- [9]. Aparna, K. H., Vidhya Subramanian, Kasirajan, M., Vijay Prakash, G., Chakravarthy, V. S. and Madhvanath, S. "Online Handwriting Recognition for Tamil", International Workshop on Frontiers in Handwriting Recognition, pp. 438-443, 2004.