International Journal Of Computer Science And Applications Vol. 1, No. 3, December 2008
ISSN 0974-1003

An Overview of Character Recognition Focused On Off-line Handwriting

Ms V. A. Gaikwad
Lecturer, E & TC Dept. GSMCOE, Pune
Vaishali.gaikwad6@ gmail.com
Contact No.9970923290

Dr. D.S.Bormane
Principal, RSCOE, Tathawade, Pune
bdattatray@yahoo.com

ABSTRACT
Character Recognition (CR) 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. This material serves as a guide and update for the readers, working in the Character Recognition area. First, an overview of CR systems and their evolution over time is presented. Then, the available CR techniques with their superiorities and weaknesses are reviewed. Finally, the current status of CR is discussed.

Index Terms--Character Recognition, Off-line Handwriting Recognition, Segmentation, Feature Extraction, Training and Recognition

1 INTRODUCTION
Machine simulation of human functions has been a very challenging research field since the advent of digital computers. 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. In this overview, Character Recognition (CR) is used as an umbrella term, which covers all types of machine recognition of characters in various application domains. The overview serves as emphasizing the methods required for the increasing needs in newly emerging areas, such as development of electronic libraries, multimedia databases and systems which require handwriting data entry. The study investigates the direction of the CR research, analyzing the limitations of methodologies for the systems, which can be classified based upon two major criteria: the data acquisition process (on-line or off-line) and the text type (machine-printed or hand- Manuscript) received. No matter which class the problem belongs, in general there are five major stages in the CR problem:

- Pre-processing
- Segmentation
- Representation
- Training and recognition
- Post processing.

The paper is arranged to review the CR methodologies with respect to the stages of the CR systems. Although the off-line and on-line character recognition techniques have different approaches, they share a lot of common problems and solutions. Off-line handwritten character recognition is selected as a focus of attention in this article. The comprehensive survey on off-line and on-line handwriting recognition in the book [5], which covers the Optical Character Recognition methodologies, can be taken as good starting points to reach the recent studies in various types and applications of the CR problem.

2 HISTORY
Writing, which 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 not only because of its challenge on simulation of human reading, but also, because it provides efficient applications such as the automatic processing of bulk amount of papers, transferring data into mahines and web interface to paper documents. Historically, CR systems have evolved in three ages:

1900-1980 Early ages-- The history of character Recognition can be traced as early as 1900, when the Russian Scientist Trying attempted to develop an aid for visually handicapped [2]. The first character recognizers appeared in the middle of the 1940s with the development of the digital computers However, some studies on Japanese, Chinese, Hebrew, Indian, Cyrillic, Greek and Arabic characters and numerals in both machine-printed and handwritten cases were also initiated [6], [7]. The commercial character recognizers were available in 1950s, when electronic tablets capturing the x-y coordinate data of pen-tip movement was first introduced. This innovation enabled the researchers to work on the on-line handwriting recognition problem.

1980-1990 Developments-- The studies until 1980 suffered from the lack of powerful computer hardware and data acquisition devices. 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. Historical review of CR research and development during this period can be found in [4] for off-line and on-line case.

After 1990 Advancements: The real progress on CR systems is achieved during this period. In the early nineties, Image Processing and Pattern Recognition techniques are efficiently combined with the Artificial Intelligence methodologies. Nowadays, in addition to the more powerful computers and more accurate electronic equipments such as scanners, cameras and electronic tablets. we have efficient, modern use of methodologies such as Neural Networks, Hidden Markov Models, Fuzzy Set Reasoning and Natural Language Processing [3], [1].

3 3. CHARACTER RECOGNITION (CR) SYSTEMS

In this section, we classify the available CR systems according to the data acquisition techniques and the text type as follows:

3.1 Systems Classified According to the Data Acquisition Techniques

The progress in CR methodologies evolved in two categories according to the mode of data acquisition, as online and off-line 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 on-line handwriting recognition problem has a number of distinguishing features like

- 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.
- Segmentation is easy.

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

1. The writer requires special equipment, which is not as comfortable as pen and paper.

2. It cannot be applied to documents printed or written on papers.

3. The available systems are slow and recognition rates are low for handwriting that is not neat.

Applications of on-line character recognition systems include small hand-held devices like pen based computers, educational software for teaching handwriting and signature verifiers are the examples of popular tools utilizing the on- line character recognition techniques. 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 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 major advantage of the off-line recognizers is to allow the previously written and printed texts to be processed and recognized.

Some applications of the off-line recognition are large-scale data processing such as postal address reading, check sorting, office automation for text entry, automatic inspection and identification. Off-line character recognition is a very important tool for creation of the electronic libraries.

3.2 Systems Classified According to the Text Type

Considering the text type, hand-written and machine-printed character recognition systems are two main areas of interest in the CR field. Machine-printed text includes the materials such as books, newspapers, magazines, documents and various writing units in the video or still image. When the documents are generated on a high quality paper with modern printing technologies, the available systems yield as well as 99% recognition accuracy. However, the recognition rates of the commercially available products are very much dependent on the age of the documents, quality of paper and ink, which may result in significant data acquisition noise.

4 METHODOLOGIES OF CR SYSTEMS

In this section, we focus on the methodologies of CR systems, emphasizing the off-line handwriting recognition problem. A hierarchical approach for most of the systems would be from pixel to text, as follows:

Pixel →Feature →Character →Sub –word →Word → Meaningful text

This bottom up approach varies a great deal, depending upon the type of the CR system and the methodology used.

4.1 Pre-processing

The raw data, depending on the data acquisition type, is subjected to a number of preliminary processing steps to make it usable in the stages of character analysis. Preprocessing aims to produce data that are easy for the CR systems to operate accurately. The main objectives of preprocessing are: Noise reduction, Normalization of the data

4.2 Segmentation

Segmentation is an important stage, because the extent one can reach in separation of words, lines or characters directly affects the recognition rate of the script. There are two types of segmentation:

- External Segmentation
- Internal Segmentation
- Explicit Segmentation
- Implicit Segmentation

4.3 Representation

Image representation plays one of the most important roles in a recognition system. In the simplest case, gray-level or binary images are fed to a recognizer. However, in most of the recognition systems, in order to avoid extra complexity and to increase the accuracy of the algorithms, a more compact and characteristic representation is required. A good survey on feature extraction methods for character recognition can be found in [8]. Hundreds of document image representation methods are categorized in three major groups as:

- Global Transformation and Series Expansion

- Statistical Representation
- Geometrical and Topological Representation

4.4 Training and Recognition Techniques

CR systems extensively use the methodologies of pattern recognition, which assigns an unknown sample into a predefined class. Numerous techniques for CR can be investigated in four general approaches of Pattern Recognition, as suggested as

- Template Matching,
- Statistical Techniques,
- Structural Techniques,
- Neural Networks.

4.4.1 TEMPLATE MATCHING

The simplest way of character recognition is based on matching the stored prototypes against the character or word to be recognized (group of pixels, shapes, curvature etc.). Matching techniques can be studied in three classes:

- Direct Matching
- Deformable Templates and Elastic Matching
- Relaxation Matching

4.4.2 STATISTICAL TECHNIQUES

The major statistical approaches, applied in the CR field are the followings:

- Non-parametric Recognition:
- Parametric Recognition:
- Clustering Analysis:
- Hidden Markov Modeling (HMM):

Hidden Markov Models are the most widely and successfully used technique for handwritten character recognition problem. There are two basic approaches to CR systems using HMM:

- Model Discriminant HMM
- Path Discriminant HMM

4.4.3 STRUCTURAL TECHNIQUES

These patterns are used to describe and classify the characters in the CR systems. The characters are represented as the union of the structural primitives. The following structural methods are applied to the CR problems:

- Grammatical Methods
- Graphical Methods

4.4.4 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. Because of its adaptive nature, it can adapt to changes in the data and learn the characteristics of input signal. A neural network contains many nodes. Neural network architectures can be classified into two major groups, namely, feed-forward and feedback (recurrent) networks. The most common neural networks used in the CR systems are the multilayer perceptron of the feed forward networks. Some of the connections are variable and can be modified by learning.

4.5 Post Processing:

It is well known that humans read by context up to 60% for careless handwriting. While preprocessing tries to “clean” the document in a certain sense, it may remove important information, since the context information is not available at this stage. The review of the recent CR research indicates minor improvements, when only shape recognition of the character is considered. Therefore, the incorporation of context and shape information in all the stages of CR systems is necessary for meaningful improvements in recognition rates. This is done in the post processing stage with a feedback to the early stages of CR. The simplest way of incorporating the context information is the utilization of a dictionary for correcting the minor mistakes of the CR systems.

5 DISCUSSION

In this study, we have overviewed the main approaches used in the CR field. Our attempt was to bring out the present status of CR research. Although each of the methods summarized above have their own superiorities and drawbacks, the presented recognition results of different methods seem very successful. However, it is very difficult to make a judgment about the success of the results of recognition methods, especially in terms of recognition rates, because of different databases, constraints and sample spaces. For texts which are handwritten under poor conditions or for free style handwriting, there is still an intensive need in al

most all the stages of CR research. A popular application area is number digit or limited vocabulary form (bank checks, envelopes and forms designed for specific applications) recognition. The best OCR packages in the market use combined techniques based on neural networks for machine- printed characters.

6 REFERENCES

[1] J. Hu, S. G. Lim, M.K. Brown, “Writer Independent On-line Handwriting Recognition Using an HMM Approach”, Pattern Recognition, vol.33, no.1, 2000.

[2] J. Mantas, “An Overview of Character Recognition Methodologies”, Pattern Recognition, vol.19, no.6, pp. 425-430, 1986.

[3] A. Meyer, “Pen Computing: A Technology Overview and A Vision”, SIGCHI Bulletin, vol.27, no.3, pp.46-90, 1995.

[4] S. Mori, C. Y. Suen, K. Yamamoto, “Historical Review of OCR Research and Development”, IEEE-Proceedings, vol.80, no.7, pp.1029- 1057, 1992.

[5] S. Mori, H. Nishida, H. Yamada, Optical Character Recognition, Wiley, 1999.

[6] S. Mori, K. Yamamoto, M. Yasuda, “Research on Machine Recognition of Handprinted Characters”, IEEE Trans. Pattern Analysis, Machine Intelligence, vol.6, no.4, pp.386-404, 1984.

[7] C. Y. Suen, M. Berthod, S. Mori, “Automatic Recognition of Handprinted Characters - The State of the Art”, Proc. of the IEEE, vol. 68, no. 4,1980.