ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 2, February 2015

The State of the Art in Text Recognition Techniques

Rohini Salunke[1], Dipali Badhe[2], Vanita Doke[3], Yogeshwari Raykar[4], Prof. Bhushan S. Thakare[5]

BE Student, Department of Computer Engineering, Sinhgad Academy of Engineering, Pune, India[1,2,3,4]

Department of Computer Engineering, Sinhgad Academy of Engineering, Pune, India[5]

Abstract: In this paper we present a comprehensive review of text recognition techniques. Character recognition has gained lot of attention in the field of pattern recognition due to its application in various fields. In coming days, character recognition system might serve as a key factor to create paperless environment by digitizing and processing existing paper documents. We believe that the review should prove helpful in identifying and solving problems that are being faced in developing a practical system.

Keywords: Text recognition techniques, Categories , Text Recognition, Segmentation, OCR, ACR, HCR

I. INTRODUCTION

Character recognition is an art of detecting segmenting and identifying characters from image. More precisely Character recognition is process of detecting and recognizing characters from input image and converts it into ASCII or other equivalent machine editable form[1][2][3].

It contributes immensely to the advancement of automation process and improving the interface between man and machine in many applications [4].Lots of independent work is going on in Optical Character Recognition that is processing of printed/computer generated document and handwritten and manually created document processing i.e. handwritten character recognition.

II. CATEGORIES

A. Online Text Recognition

On-line handwriting recognition means that the machine recognizes the writing while the user writes. The term real time or dynamic has been used in place of online. In case of online character recognition system, character is processed while it was under creation. External factors like pressure speed of writing, stroke making etc. have great impact on online system. On-line handwriting recognition requires a transducer that captures the writing as it is written. The most common of these devices is the electronic tablet or digitizer, which typically has a resolution of 200 points/in, a sampling rate of 100 points/s, and an indication of "inking" or pen down [5].

B. Offline Text Recognition

In case of offline character recognition system, document is first generated, digitized, stored in computer and then it is processed. External factors like pressure speed of writing, stroke making etc. does not have any influence in case of offline system. Offline handwriting recognition, by contrast, is performed after the writing is completed.

III. TECHNIQUES

There are generally two techniques used for text recognition.

A. Segmentation Of Text

Several different methods of text segmentation are compared for various image data formats. The problem of correct segmentation of joined and broken characters are also considered.

B. Contextual Recognition

Two techniques for contextual recognition are considered: the Markov-based methods and dictionary look-up methods. The various techniques for storing dictionary information are compared. A discussion of the importance of the choice of the correct context is given, together with guidance on which methods are best suited to which applications.

IV. METHODS

Many recent methods have been proposed to design better feature representations and models for both. In this paper, we apply methods recently developed in machine learning–specifically, large-scale algorithms for learning the features automatically from unlabeled data–and show that they allow us to construct highly effective classifiers for both detection and recognition to be used in a high accuracy end-to-end system.

A. OCR

Optical Character Recognition [6] – [10] is a process that can convert text, present in digital image, to editable text. It allows a machine to recognize characters through optical mechanisms. The output of the OCR should ideally be same as input in formatting. The process involves some pre-processing of the image file and then acquisition of important knowledge about written text. That knowledge or data can be used to recognize characters. OCR [6] is becoming an important part of modern research based computer applications. Especially with the advent of Unicode and support of complex scripts on personal computers, the importance of this application has increased. The aim is to preserve these documents and make them fully accessible, searchable and process able in digital form.

1) Design Of OCR

Various approaches used for the design of OCR systems are discussed below:

DOI 10.17148/IJARCCE.2015.42100

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 2, February 2015

a) Matrix matching [11]: Matrix Matching converts each character into a pattern within a matrix, and then compares the pattern with an index of known characters. Its recognition is strongest on monotype and uniform single column pages.
b) Fuzzy logic [11]: Fuzzy logic is a multi-valued logic that allows intermediate values to be defined between conventional evaluations like yes/no, true/false, black/ white etc. An attempt is made to attribute a more human-like way of logical thinking in the programming of computers. Fuzzy logic is used when answers do not have a distinct true or false value and there is uncertainly involved.
c) Feature extraction [8][11]: This method defines each character by the presence or absence of key features, including height, width, density, loops, lines, stems and other character traits. Feature extraction is a perfect approach for OCR of magazines, laser print and high quality images.
d) Structural analysis [11]: Structural Analysis identifies characters by examining their sub features shape of the image, sub-vertical and horizontal histograms. Its character repair capability is great for low quality text and newsprints.
e) Neural networks [11]: This strategy simulates the way the human neural system works. It samples the pixels in each image and matches them to a known index of character pixel patterns. The ability to recognize characters through abstraction is great for fixed documents and damaged text. Neural networks are ideal for specific types of problems, such as processing stock market data or finding trends in graphical patterns.

2) Stages Of OCR

Various stages of OCR system design are given in fig 1.

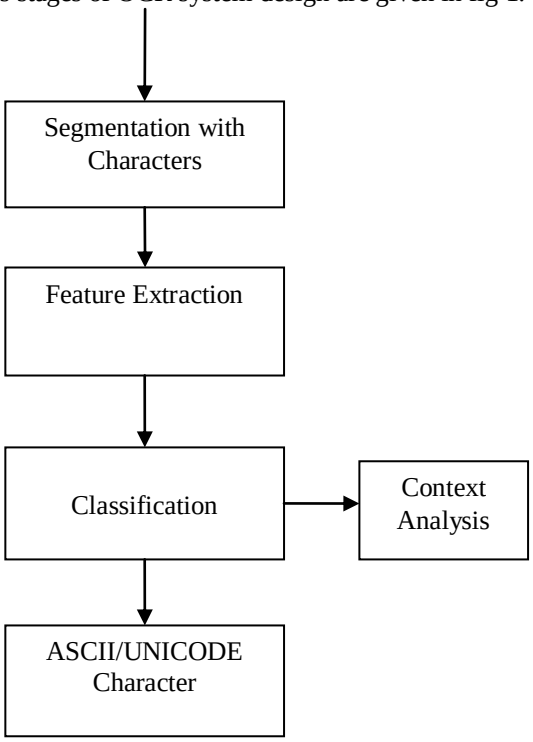

Fig 1: Stages in OCR Design (adapted from [11])

B. ACR

Arabic Character Recognition: During the past three decades, considerable research and development works have been done towards the development of an efficient Arabic optical character recognition (ACR) system. We believe that the review should prove helpful in identifying and solving problems that are being faced in developing a practical ACR system.

1) Characteristics of Arabic Text

An Arabic text is composed by placing linearly character blocks of varying sizes from right to left. The peculiar characteristic of the Arabic text is that the shape of characters may significantly vary within a word[15]. This variation depends on: the position of the character within a word and its adjacent characters. On the basis of characteristics, Arabic text can be categorized into various types.

2) Text Styles

An Arabic text can be produced in any of the following styles.

a) Unconstrained non-isolated handwritten text
b) Unconstrained isolated handwritten text
c) Constrained non-isolated handwritten text
d) Constrained isolated handwritten text
e) Unifont typewritten text: Typewritten text that involves only one font.
f) Multifont typewritten text: Typewritten text involving many fonts.

3) Factors affecting the design of ACR

We classify factors affecting the development of an *ACR* system into two classes: *random* factors and *linguistic* factors.

a) Random Factor:

Random factor**s** affect the document scanning process. Examples of random factors are: the document digitization errors, ink and dirt spattering, paper quality, the quality of writing tools, random distortions introduced by the scanning devices, etc.

b) Linguistic Factor:

Linguistic factor**s** are the intrinsic part of the Arabic language. The cardinality of the Arabic alphabet set is one of the linguistic factors that affect the design of *ACR* system. Character shape variation within a word is linguistic factor that affects the design of an *ACR* system. Writing style and character connectivity are the two major sources of shape variations. shape variation poses many problems in developing an *ACR* system for Arabic alphabet based languages.

C. HCR

Handwriting Character Recognition means that the machine recognizes the writing while the user writes. Handwriting consists of a time sequence of strokes, where a stroke is the writing from pen down to pen up. The characters of writing are usually formed in sequence, one character being completed before beginning the next, and the characters typically follow some spatial order, such as left to right. Handwritten Chinese characters are usually separated spatially, one from the other; in fact, they are

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 2, February 2015

often written in boxes [14]. Handwritten English words are normally separated spatially and are often written on lined paper. Letters within a word, however, are not usually separated spatially.

1) Characteristics of Arabic Text

Off-line handwriting recognition is performed after the writing is completed. The writing is usually captured by an optical scanning device.

On-line (real-time, dynamic) handwriting recognition is machine recognition of writing as it is being written on a table digitizer.

2) Various States

Pen down, for the normal situation of writing with pen on paper, is simply the state in which the pen is "inking," while *pen up* is the non inking state. For electronic tablets, pen down is the electronic equivalent of the state in which inking occurs. For many tablets, a micro switch in the pen tip closes when the pen is in contact with the tablet surface to indicate pen down.

3) Processing Stages

Post processing is processing of the output from shape recognition.

Preprocessing is processing of the handwriting data prior to shape recognition.

Segmentation is the machine separation of writing units from each other. External segmentation does not require recognition, while internal does.

Shape recognition is the pattern recognition of shapes of writing units.

4) Key Points

A stroke consists of the writing from pen down to pen UP.

A stroke segment is a stroke or portion of a stroke. Writing units are clearly defined units of writing, such as strokes, characters, and words. Stroke segments also qualify if clearly defined.

ACKNOWLEDGMENT

We would like to give our sincere gratitude to our guide **Mr. Bhushan Thakare** who guided us throughout, to complete this paper.

REFERENCES

[1] Kai Ding, Zhibin Liu, Lianwen Jin, Xinghua Zhu, "A Comparative study of GABOR feature and gradient feature for handwritten 17hinese character recognition", International Conference on Wavelet Analysis and Pattern Recognition, pp. 1182-1186, Beijing, China,2-4 Nov. 2007

[2] Pranob K Charles, V.Harish, M.Swathi, CH. Deepthi, "A Review on the Various Techniques used for Optical Character Recognition", International Journal of Engineering Research and Applications, Vol. 2, Issue 1, pp. 659-662, Jan-Feb 2012

[3] Om Prakash Sharma, M. K. Ghose, Krishna Bikram Shah, "An Improved Zone Based Hybrid Feature Extraction Model for Handwritten Alphabets Recognition Using Euler Number", International Journal of Soft Computing and Engineering, Vol.2, Issue 2, pp. 504-58, May 2012

[4] J. Pradeepa, E. Srinivasana, S. Himavathib, "Neural Network Based Recognition System Integrating Feature Extraction and Classification for English Handwritten", International journal of Engineering,Vol.25, No. 2, pp. 99-106, May 2012

[5] A. Amin, A. Kaced, J. Haton, and R. Mohr, "Hand written Arabic character recognition by the **I.R.A.C.** sytem," in *Proc. 5th Int. Conf. Pattern Recognition, 1980*, pp. 729-731.

[6] Dan ClaudiuCires¸an and Ueli Meier and Luca Maria Gambardella and JurgenSchmidhuber, "Convolutional Neural Network Committees for Handwritten Character Classification", 2011 International Conference on Document Analysis and Recognition, IEEE, 2011.

[7] GeorgiosVamvakas, Basilis Gatos, Stavros J. Perantonis, "Handwritten character recognition through two-stage foreground sub-sampling" ,Pattern Recognition, Volume 43, Issue 8, August 2010.

[8] Shrey Dutta, Naveen Sankaran, PramodSankar K., C.V. Jawahar, "Robust Recognition of Degraded Documents Using Character N-Grams", IEEE, 2012.

[9] Naveen Sankaran and C.V Jawahar, "Recognition of Printed Devanagari Text Using BLSTM Neural Network", IEEE, 2012.