An Introduction to the Process of Optical Character Recognition

Umal Patel[1]

[1] Department of Computer Engineering
L D College Of Engineering, Gujarat Technological University, Gujarat, India

Abstract: *This paper presents an overview of methods and techniques used for feature extraction that helps in efficient classification of the alphabets and numbers of English language. Character recognition has long been a essential area for research since years. Recognition of character is a minor work for humans, but to make a computer program that does character recognition is extremely difficult. Hence to make a machine recognize the characters and efficiently determine a pattern has been the primary concern for researchers now days This paper discusses various offline and online Optical Character Recognition Techniques (OCR).*

Keywords: OCR, online, offline, online, zoning, euler number.

1. Introduction

OCR is an approach that provides a full alphanumeric recognition of printed or handwritten characters at electronically by simply scanning them and generating into a form that can be scanned through a scanner and then the recognition engine of the OCR system interpret the images and turn images of handwritten or printed characters into ASCII data (machine-readable characters).Character recognition also popularly referred as optical character recognition (OCR) is a field of research that has immense potential in future where we want to track and locate every piece of information being exchanged. The problem with the hand written text is due to uncertainties such as variation in calligraphy over period of time, similarity in text, variation in styles of writing [3] The character recognition system helps in making the communication between a human and a computer easy.[4] The character recognition is basically classified into two types: offline handwritten text recognition, online handwritten text recognition. Offline means the text written on the plain paper or sheet and then the writing is usually captured optically by a scanner and the completed writing is available as an image. Online means the text written on any digital devices such as tablets using stylus i.e. the two dimensional coordinates of successive points are represented as a function of time and the order of strokes made by the writer are also available.[6]

2. Applications of optical character recognition

The area of OCR is becoming an integral part of document scanners, and is used in many applications such as postal processing, script recognition, banking, security (i.e. passport authentication) and language identification, document reading, mail sorting, signature verification, writer identification., license plate recognition system, smart card processing system, automatic data entry, bank cheque /DD processing, money counting machine, postal automation, address and zip code recognition etc many organizations are depending on OCR systems to eliminate the human interactions for better performance and efficiency [2,4,6,7].

3. Potential problem areas for OCR

1. The same characters differ in sizes, shapes and styles from person to person and even from time to time with the same person. The source of confusion is the high level of abstraction: there are thousands styles of type in common use plus variations in calligraphy and a character recognition program must recognize most of these.
2. Like any image, visual characters are subject to spoilage due to noise. Some images containing characters are already blurred or not clear which makes them difficult to process. Noise consists of random changes to a pattern, particularly near the edges. A character with much noise may be interpreted as a completely different character by a computer program.
3. There are no hard-and-fast rules that define the appearance of a visual character. Hence rules need to be heuristically deduced from the samples.

4. Phases of OCR
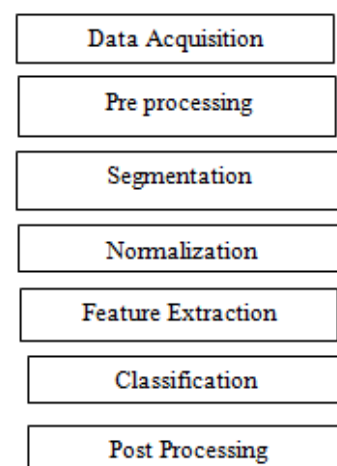
1. Data Acquisition

Most Important initial phase in OCR is to gather the image from either device sensor like PDA or tablets in case on online recognition or getting the images containing characters directly for offline recognition.

In Image acquisition, the recognition system acquires a scanned image as an input image. The image should have a specific format such as JPEG, BMP etc. This image is acquired through a scanner, digital camera or any other suitable digital input device. Data samples for the experiment have been collected from different individuals [9].

2. Pre Processing

The goal of pre-processing is to simplify the pattern recognition problem without missing any vital information. It reduces the noises and inconsistent data. It enhances the image and prepares it for the next steps [3].

Preprocessing is the preliminary step which transforms the data into a format that will be more easily and effectively processed. Therefore, the main task in preprocessing the captured data is to decrease the variation that causes a reduction in the recognition rate and increases the complexities, as for example, preprocessing of the input raw stroke of characters is crucial for the success of efficient character recognition systems. Thus, preprocessing is an essential stage prior to feature extraction since it controls the suitability of the results for the successive stages [2].

Preprocessing can be done through various ways Binarization, Noise reduction, Stroke width normalization, Skew correction, Slant removal, Filtering, Morphological Operations, Noise Modelling, Skew Normalization, Size Normalization, Contour Smoothing, Compression, Thresholding, Thinning etc
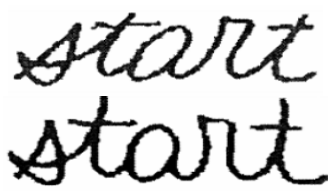

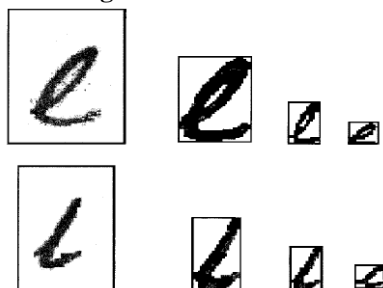**Figure 1:** Slant Removal

Figure 2: Normalization of 'e' and 'l' as in [9]

3. Segmentation

Segmentation is an integral part of any text based recognition system. It assures efficiency of classification and recognition. Accuracy of character recognition heavily depends upon segmentation phase.
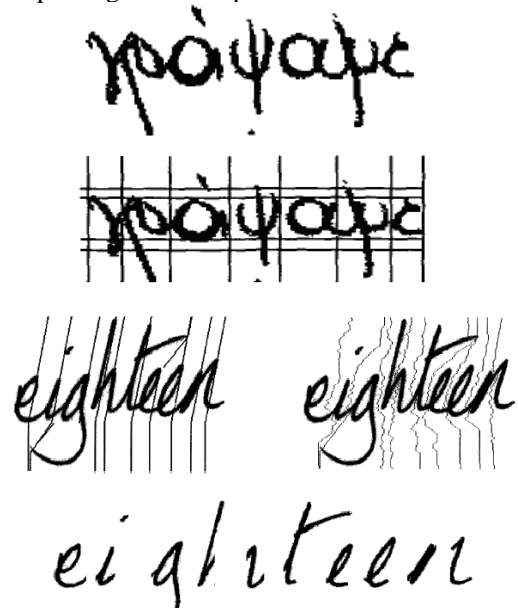

Figure 3: Segmentation [9]

4. Normalization

The results of segmentation process provides isolated characters which are ready to pass through feature extraction stage, thus the isolated characters are reduced to a specific size depending on the methods used. The segmentation process essentially renders the image in the form of m*n matrix. These matrices are then generally normalized by reducing the size and removing the redundant information from the image without losing any important information.

5. Feature Extraction

Feature extraction is the process of extracting the relevant features from objects/alphabets to form a feature vectors. These feature vectors is then used by classifiers to recognize the input unit with target output unit. It becomes easier for the classifier to classify between different classes by looking at these features as it allows fairly easy to distinguish.

Feature extraction is also defined as extracting the raw data the information which is most relevant for classification purposes in the sense of minimizing the pattern variability.[1]
Due to the nature of handwriting with its high degree of variability and imprecision obtaining these features, is a difficult task. Feature extraction methods are based on 3 types of features:

- • Statistical
- • Structural
- • Global transformations and moments

Statistical Features includes:

1. Zoning

The character image is divided into NxM zones. From each zone features are extracted to form the feature vector. The

goal of zoning is to obtain the local characteristics instead of global characteristics.
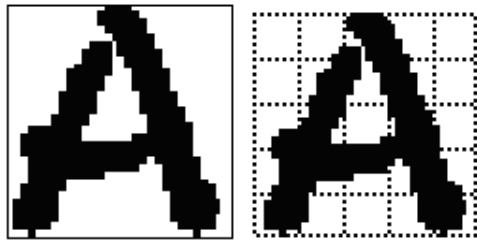

Figure 4: zoning

After dividing the character into different zones you can compare the density or direction features of it and classify each one of them.

2. Projection Histograms

The basic idea behind using projections is that character images, which are 2-D signals, can be represented as 1-D signal. These features, although independent to noise and deformation, depend on rotation. Projection histograms count the number of pixels in each column and row of a character image. Projection histograms can separate characters such as "m" and "n".
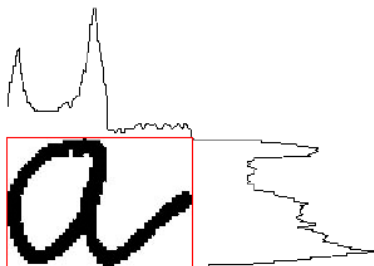

Figure 5: Projection Histogram

3. Profiles

The profile counts the number of pixels (distance) between the bounding box of the character image and the edge of the character. The profiles describe well the external shapes of characters and allow distinguishing between a great number of letters, such as "p" and "q".

Figure 6: Profiling

4. Structural features:

Structural features are based on topological and geometrical properties of the character, such as aspect ratio, cross points, loops, branch points, strokes and their directions, inflection between two points, horizontal curves at top or bottom, etc.

Global Transformations-Moments:

The Fourier Transform (FT) of the contour of the image is calculated. Since the first n coefficients of the FT can be used in order to reconstruct the contour, then these n coefficients are considered to be a n-dimensional feature vector that represents the character.
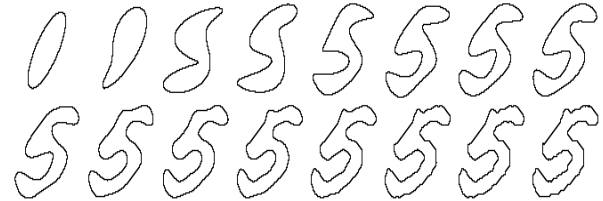

Figure 7: Contouring

6 Classification

The results Classification is the last stage where we train the neural net using the feature vectors obtained during feature extraction method against the required targets. To optimize the whole recognition process, several combination methods of multilayer perceptron have been devised. E.g.: k-Nearest Neighbour (k-NN), Bayes Classifier, Neural Networks (NN), Hidden Markov Models (HMM), Support Vector Machines (SVM), etc there is no such thing as the "best classifier". The use of classifier depends on many factors, such as available training set, number of free parameters etc.

7. Post Processing

The goal of post processing is the incorporation of context and shape information in all the stages of OCR systems is necessary for meaningful improvements in recognition rates.

5. Conclusion

The character recognition methods have been introduced and developed over the years. In this paper, I have tried to explain the overview of the whole OCR process and the methods related to it. Many researchers try to hybrid two or more different methods and compare the results for efficiency but again this will be application specific and parameter specific. OCR has been implemented in various countries for recognizing different languages as well.

References

[1] Oivind Due Trier, Anil K. Jain, Torfinn Taxt, "Feature Extraction Methods for Character Recognition-A Survey", July 1995
[2] Yasser Alginahi, Taibah University Kingdom of Saudi Arabia, "Preprocessing Techniques in Character Recognition"
[3] Om Prakash Sharma, M. K. Ghose, Krishna Bikram Shah, Benoy Kumar Thakur, "Recent Trends and Tools for Feature Extraction
[4] in OCR Technology", International Journal of Soft Computing and Engineering (IJSCE)
[5] ISSN: 2231-2307, Volume-2, Issue-6, January 2013

[6] Suruchi G. Dedgaonkar, Anjali A. Chandavale, Ashok M. Sapkal, "Survey of Methods for Character Recognition", International Journal of Engineering and Innovative Technology (IJEIT) Volume 1, Issue 5, May 2012

[7] Mohanad Alata, Mohammad Al-Shabi "TEXT DETECTION AND CHARACTER

[8] RECOGNITION USING FUZZY IMAGE PROCESSING", Journal of ELECTRICAL ENGINEERING, VOL. 57, NO. 5, 2006, 258–267

[9] Rejean Plamondon, Fellow, IEEE and Sargur N. Shrihari, Fellow, IEEE, "On-line and Off-line Handwriting Recognition: A comprehensive Survey", IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL 22, NO. 1 JANUARY 2000

[10] 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 (IJSCE) ISSN: 2231-2307, Volume-2, Issue-2, May 2012

[11] J. Pradeepa,, E. Srinivasan, S. Himavathi, "Neural Network Based Recognition System Integrating Feature Extraction and Classification for English Handwritten", IJE TRANSACTIONS B: Applications Vol. 25, No. 2, (May 2012) 99-106

[12] Nafiz Arica and Fatos T. Yarman-Vural, "An Overview of Character Recognition Focused on

[13] Off-Line Handwriting", IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART C: APPLICATIONS AND REVIEWS, VOL. 31, NO. 2, MAY 2001

Author Profile

Umal Patel, pursuing her Master Degree in Computer Science & Technology from Gujarat Technological University (L D College of Eng., Ahmedabad), received her Bachelor Degree in Computer Engg. from Gujarat University in 2008. Presently working as Assistant Professor in Department of MCA, L.J Institute of Technology. Earlier he has served as Software Test Engineer in Lodestone Software Services since June 2008. Her area of interest is Compilers and Image Processing.