ORIENTAL JOURNAL OF
COMPUTER SCIENCE & TECHNOLOGY
An International Open Free Access, Peer Reviewed Research Journal
Published By: **Oriental Scientific Publishing Co., India.**
www.computerscijournal.org

ISSN: 0974-6471
June 2012,
Vol. 5, No. (1):
Pgs. 41-47

An Overview of Character Recognition Systems

K.B. GEETHA

Pooja Bhagavat Memorial Mahajana PG Centre, DoS in CS,
KRS Road, Metagalli, Mysore - 570 016 (India).
E-mail: kbgeethaumesh@yahoo.co.in

(Received: February 12, 2012; Accepted: June 04, 2012)

ABSTRACT

Character recognition is the processing by machine of text based input patterns to produce some meaningful output. Character recognition lies at the core of the discipline of pattern recognition where the aim is to represent a sequence of characters taken from alphabets. The advancements in pattern recognition has accelerated recently due to the many emerging applications which are not only challenging, but also computationally more demanding, such as Character Recognition, Document Classification, Computer Vision, Data mining, Shape recognition and Biometric Authentication etc. The area of Optical Character Recognition (OCR) is becoming an integral part of document scanners and is used in many applications like postal processing, script recognition, banking security. Where OCR is a part of the character recognition. The research in this area has been ongoing for over half a century and the outcomes have been astounding with successful recognition rates for printed characters 99%, with significant improvements in performance for handwritten cursive character recognition where recognition rates have exceeded the 90%. Nowadays, many organizations are depending on OCR systems to eliminate the human interactions for better performance and efficiency. Because of this, it is necessary to know about the character recognition system. This paper helps to beginners, by presenting an overview of the character recognition system and its functional components.

Keywords: Pattern Recognition, Character Recognition System (CRS), Optical Recognition System (OCR).

INTRODUCTION

Character Recognition technology has a long history of research. Various algorithms, from classical template matching and multiple similarities to the latest neural network theory have been proposed. The basic theory of pattern recognition technology, including character recognition technology, is still in the research and development stages. The field of Pattern recognition is a multidisciplinary field which forms the foundation of other fields, such as Image Processing, Machine Vision and Artificial Intelligence.

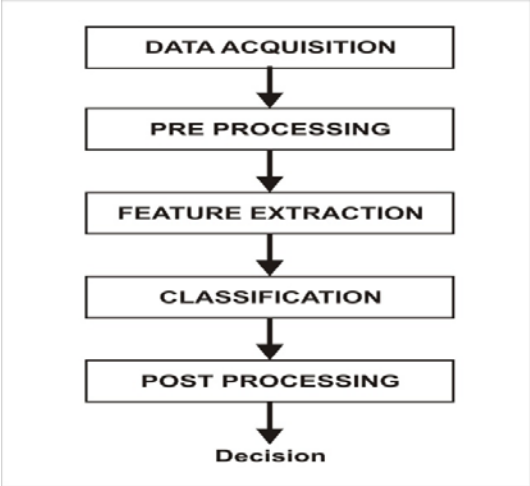
Character Recognition basically converting scanned images in to text document can enable manipulation through word processing applications. Optical Character Recognition has gained a momentum since the need for digitizing or converting scanned images of machine printed or hand written text (numerals, letters, and

symbols), in to a format recognized by computers (such as ASCII). OCR has been extensively used as the basic application of different learning methods in machine learning literature[11]. Handwriting recognition is the task of transforming a language re-presented in its own spatial form of graphical marks into a symbolic representation[12]. Handwriting recognition inherited a number of technologies from optical character recognition (OCR). The main difference between handwritten and typewritten characters is in the variations that come with handwriting. It is also worth noticing that OCR deals with off-line recognition while handwriting recognition may be required for both on-line and off-line signals.

Character Recognition uses computer software to translate paper documents into electronic word documents. There are different types of character recognition software exist, with some having the ability to scan handwritten documents.

Three major types of character recognition software currently exist are

- **OCR** known as Optical Character Recognition uses a computer scanner. It is a computer program designed to convert scanned or digital images of handwritten or typewritten text into machine-editable text, or to translate pictures of characters into a standard encoding scheme representing them.
- **ICR** known as Intelligent Character Recognition is used by business and government agencies to read standard forms. The application of this software is limited to specific fields.
- **Natural Handwriting Recognition** reads handwriting, usually by comparing whole words, based upon a software database.



Functional Components of Character Recognition System

Functional components of character Recognition system

Character recognition system contains numerous functional components including: Data acquisition, Pre-processing, Feature extraction, Classification and Post-processing.

i. Data Acquisition

To Character Recognition System (CRS), input may come from on-line or off-line devices. On-line devices are stylus based. Off-line devices include scanners and hand-held types. Typical data acquisition devices for off-line and on-line recognition are scanners and digitizing tablets, respectively. Due to the lack of temporal information, off-line handwriting recognition is considered more difficult than on-line.

ii. Pre-processing

The most crucial aspect in character recognition is the preprocessing, which is necessary to modify the data either to correct deficiencies in the data acquisition process due to limitations of the capturing device sensor, or to prepare the data for subsequent activities later in the description or classification stage. Data preprocessing describes any type of processing performed on raw data to prepare it for another processing procedure. Hence, preprocessing is the preliminary step which transforms the data into a format that will be more easily and effectively processed in further steps. Therefore, the main task of preprocessing is to decrease the variations in the captured data which causes a reduction in the recognition rate and increase the complexities. Thus, preprocessing is an essential phase prior to feature extraction, since it controls the suitability of

the results for the successive phases. The components in a CRS are in a pipeline fashion, each stage depends on the success of the previous stage in order to produce optimal results.

The main objective of the preprocessing stage is to normalize and remove the factors, which will complicate the classification and reduces the recognition rate. Hence preprocessing includes the functions: image enhancement, noise reduction, image thresholding, skew detection/ correction, Skeletonisation, character segmentation and normalization.

- ***Image enhancement techniques***

Image enhancement improves the quality of images for human perception by removing noise, reducing blurring, increasing contrast and providing more detail. Some of the techniques used in image enhancement are spatial image filtering operations, Mask processing, local thresholding.

- ***Noise Reduction***

Noise is a random error in pixel value, usually introduced as a result of reproduction, digitalization and transmission of the original image. Noise is an important factor that influences the image quality, which is mainly produced in the processes of image acquirement and transmission. The presence of noise can cost the efficiency of the character recognition system. Noise cannot always be totally eliminated, but smoothing is a widely used procedure for replacing the value of a pixel by the average of the values of the pixels around the original pixel.

- ***Image thresholding***

Image thresholding is the process of separating the information (objects) of an image from its background, hence, thresholding is usually applied to grey-level or color document scanned images. Thresholding can be categorized into two main categories: global and local. Global thresholding methods choose one threshold value for the entire document image, which is often based on the estimation of the background level from the intensity histogram of the image; hence, it is considered a point processing operation. Global thresholding methods are used to automatically reduce a grey-level image to a binary image. The images applied to such methods are assumed to have two classes of pixels (foreground and background). The purpose of a global thresholding method is to automatically specify a threshold value, *T*, where the pixel values below it are considered foreground and the values above are background. A simple method would be to choose the mean or median value of all the pixels in the input image, the mean or median will work well as the threshold, however, this will generally not be the case especially if the pixels are not uniformly distributed in an image. A more sophisticated approach might be to create a histogram of the image pixel intensities and use the valley point (minimum) as the threshold. The histogram approach assumes that there is some average value for the background and object pixels, but that the actual pixel values have some variation around these average values. However, this may be computationally expensive, and image histograms may not have clearly defined valley points, often making the selection of an accurate threshold difficult. One method that is relatively simple and does not require much specific knowledge of the image is the iterative method.

Local thresholding techniques are used with document images having non-uniform background illumination or complex backgrounds, such as watermarks found in security documents if the global thresholding methods fail to separate the foreground from the background. This is due to the fact that the histogram of such images provides more than two peaks making it difficult for a global thresholding technique to separate the objects from the background, thus; local thresholding methods are the solution. The local thresholding techniques developed in the literature are mainly for specific applications and most of the time they do not perform well in different applications. The results could be over thresholding or under thresholding depending on the contrast and illumination.

- ***Skeletonisation***

There are two basic techniques for producing the skeleton of an object: *basic thinning and medial axis transforms.* Thinning is a morphological operation that is used to remove selected foreground pixels from binary images, somewhat like erosion or opening. Thinning is a

data reduction process that erodes an object until it is one-pixel wide, producing a skeleton of the object making it easier to recognize objects such as characters. Figure shows how thinning the character **'E'** produces the skinny shape of the character. Thinning is normally only applied to binary images, and produces another binary image as output. Thinning erodes an object over and over again (without breaking it) until it is one-pixel wide. On the other hand, the medial axis transform finds the points in an object that form lines down its center[#]. The medial axis transform is similar to measuring the Euclidean distance of any pixel in an object to the edge of the object, hence, it consists of all points in an object that are minimally distant to more than one edge of the object[*].

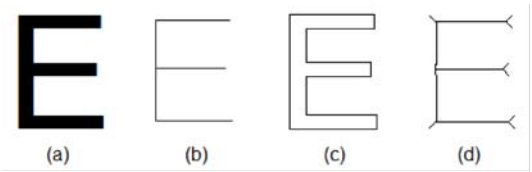


Fig. 2: (a) Original Image (b) Medial Axis Transform (c) Outline (d) Thinning

- ***Skew detection/correction***

Aligning the paper document with the co-ordinate system of the scanner is essential and called as skew correction. Due to the possibility of rotation of the input image and the sensitivity of many document image analysis methods to rotation of the image, document skew should be corrected. There are around twenty five different methods for document image skew detection. The methods are based on Hough transform analysis, Projection profile, feature point distribution and orientation sensitive feature analysis. The techniques reported upto 0.1 degrees accuracy, so there is a strong need for further work in this area.

- ***Segmentation and Normalization***

Character segmentation is considered one of the main steps in preprocessing. Characters can be written cursively, where characters are connected together or may also overlap. For a CRS that is required to identify individual character there is a need to identify where a character starts and ends. This is what essentially what segmentation aims to do. There are various methods for tackling the segmentation problem:

Ø Pre-segmentation means characters that arrive already separated from each other. This is normally the case when the text is printed or when writer is required to write the characters in boxes or without connecting them together.

Ø Finding Gaps; to find out the gaps between the letters or, at least the connecting lines. Gap based techniques are stograms, bounding boxes, run-length and convex hulls. All these techniques use geometric relationships between the various components of the text.

The result from the character segmentation stage provides isolated characters which are ready to be passed into the feature extraction stage; therefore, the isolated characters are normalized into a specific size, decided empirically or experimentally depending on the application and the feature extraction or classification techniques used, then features are extracted from all characters with the same size in order to provide data uniformity.

iii. Feature extraction

Pattern recognition techniques for character recognition have been divided into two major categories:

i. Template based techniques
ii. Feature based techniques

Template based approaches aim to create a probabilistic template of each character model from the training data. During testing, the unknown pattern is superimposed directly on the ideal template pattern and degree of correlation is used to decide about the classification. To enhance the accuracy of template based results, it uses feature based approach.

Feature based approaches extract feature vectors from training samples and aim to create a class-model out of all vectors of a given class. The challenging is to derive those features which are different from other classes. The feature based approaches can be of two types, spatial domain

and transform domain approaches. Spatial domain approaches derive features directly from the pixel representation of the pattern. In a transform domain technique, the patter image is first transformed into another space using, for example: Fourier, Cosine, Slant or Wavelet transform and useful features are derived from the transformed images.

Feature extraction is more important for the proper classification. Feature extraction is one of the two most basic functions of a CRS. It involves measuring those features of the input pattern that are relevant to classification. After Feature extraction, the pattern is represented by the set of extracted features. However, only those features which are of possible relevance to classification need to be considered.

Various types to extract features are:

- Histograms – horizontal or vertical
- Curvature information
- Topological features
- Parameters of polynomial curve fitting functions
- Contour information

Feature extraction maps the whole of each input pattern from its original spatial system of co-ordinates onto a single point in a feature space. The main goal of feature extraction is to map input patterns onto points in feature space, and then the purpose of classification is to assign each point in the space with a correct label.

Classification

Once an input pattern is mapped onto a point in feature space, the next step is to label each of the points. Classification techniques includes

- Rule-based systems
- Decision Trees
- Clustering techniques

Rule-based systems uses a set of If-Then rules to decide the class of a pattern based on how well the conditions in the If-part fit the pattern. In rule based systems, it is possible for two rules to be applicable to the same input pattern. This causes conflict, and hence calls for conflict-resolution machinery.

Decision Trees viewed as a tree data structure used for decision making. A tree has a single entry point at its top, and any number of single class leaf nodes at its bottom. Once a decision tree is constructed, it can be used to classify new patterns.

Clustering techniques basically attempts to look for points in feature space that are close to each other and place them in the same class. One way of doing this is to arbitrarily assign a class to each point in a set of unclassified points, then find the centre of each group of similarly classed points. Hence, each point is re-assigned to the class of the centre-point closest to it. This is followed by a re-computation of the centre points of the various classes, and so on the process repeats until no more re-assignment is necessary.

Apart from these techniques, there is a another technique which do not explicitly derive features from the patterns. During training, after normalization the system adjusts its parameters to minimize the misclassifications. The system, thus trained is used for classifying unknown patterns. One of the most popular method is Artificial Neural Network (ANN). The ANN is an information-processing technology inspired by the way the human brain processes information. ANN's are collection of mathematical models that represent some of the observed properties of biological nervous systems and draw on the analogies of adaptive biological learning. The ANN consists of a large number of highly interconnected processing elements or nodes that are tied together with weighted connections. Learning in biological system involves adjustments to the synaptic connection that exist between the neuron, similarly in ANN also. Learning typically occurs by example through training, where training algorithm adjusts the link weights. The link weights store the knowledge necessary to solve specific problems.

Post-processing

Post-processing covers verification, action execution and adaptation. The goal of verification is to increase the level of confidence in the classification made. This is done in various ways. One way is to use a database of letter combinations to check that sequence of letters

recognized does not contain impossible combinations. Alternatively, a word dictionary could be used to check that a certain string of characters constitutes a valid word. This method fails when the word is not in dictionary.

Some advanced CRS's alter their own weights (in ANN) in an act of adaptation to reduce the gap between expected and actual performance, in an effort to improve future performance.

CONCLUSION

The problem of character recognition is a subset of pattern recognition in that it is confined to text based patterns. The aim of all CRS is to automatically extract some meaning out of some text based input. There are many types of CRS's. Some read cheques and some other recognizes printed words from a scanned image. However, all CRS's may contain five functional components.

The five functional components of a CRS are: Data acquisition, Pre-processing, Feature extraction, Classification and Post-processing. Not all CRS's have all these parts and some have additional components. But almost all CRS extract features, measures the features and classifies the input. Regardless of the techniques used, all CRS faces problem in pre-processing and adaptation. Even though many methods and techniques have been developed for preprocessing there are still problems that are not solved completely and more investigations need to be carried out in order to provide solutions. Segmentation is the biggest problem in pre-processing. It is quite difficult to find where character starts and finishes, if characters overlap, words overlap and unwanted information (noise) overlap.

The other big problem in Character Recognition is adaptation, especially in the absence of direct corrective feedback. This means that the learning would have to be unsupervised, and hence uncertain. This is because the machine will have to required to decide for itself where and what error in recognition has occurred. Today, there are several software which recognize a wide variety of fonts, but handwriting and script fonts that mimic handwriting are still problematic. Future research aims at new applications such as online character recognition used in mobile devices, extraction of text from video images, extraction of information from security documents and processing of historical documents. The objective of such research is to guarantee the accuracy and security of information extraction in real time applications.

REFERENCES

1. S.Di. Zenzo *et al*. Optical recognition of hand printed charaters of any size, position and orientation", *IBM journal of Research & Devolopment.* **36**(3) (1992).
2. Sabri A. Mahmoud " Skeletonization of Arabic Characters using Clustering based Skeletonisation Algorithm" in PR, **24**(5): 453-464 (1991).
3. A.C. Downton & S.Impedovo. "Progress in Hand-Writing Recognition", Feature Extraction chapter, published by World Scientific (1997).
4. J J Hull "Language Level Syntactic and Semantic Constraints Applied to Visual Word Recognition in Fundamentals of Handwriting Recognition", by Sebastiano Impedova, published by Springer-Verlag (1994).
5. T Y Zhyang & C Y Suen " AFast Parallel Algorithm for thinning Digital Patterns", in Communications of the *ACM*, **27**(3): (1985).
6. Mudit Agrawal "Re-targetable OCR with intelligent character segmentation"
7. Nawwaf N. Kharma & Rabab K. Ward "Character Recognition Systems for the Non-expert"
8. Yasses Alginahi "Preprocessing Techniques in Character Recognition"

9. Rafael C. Gonzalez and Richard E. Wood "Digital Image Processing" II edition, Pearson education, Low price edition.
10. R. Jagadeesh Kannan and R.prabhakar "Off-Line Cursive Handwritten Tamil Character Recognition"
11. Y. Lecun, L. Bottou, Y. Bengio, P. Haffner, "Gradient based learning applied to document recognition", Proceedings of the IEEE, vol. 86, no.11, IEEE, pp. 2278- 2324, USA , 1998
12. R. Plamondon and S. Srihari. On-line and Offline Handwriting Recognition: A Comprehensive Survey. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, **22**(1): 63–84, (2000).