Journal of Information and Operations Management
ISSN: 0976–7754 & E-ISSN: 0976–7762 , Volume 3, Issue 1, 2012, pp-196-198
Available online at http://www.bioinfo.in/contents.php?id=55

INFORMATION RETRIEVAL USING ARTIFICIAL INTELLIGENCE AND FUZZY LOGIC FOR HAND WRITTEN DOCUMENTS THROUGH OPTICAL CHARACTER RECOGNITION (OCR)

PANDEY M.K.[1] AND NANDAN SINGH DASILA[2]

AIMCA, Haldwani, India
*Corresponding Author: Email- [1]mkpbsb@yahoo.com, [2]nandansinghdasila@gmail.com

Received: December 12, 2011; Accepted: January 15, 2012

Abstract- Information retrieval systems have gained considerable momentum in the last few years. With the advent of computers, it became possible to store large amounts of information and finding useful information from such collections became a necessity. There is an urgent need for technologies that will allow efficient and effective processing of huge datasets. There is a great demand for efficient and effective ways to organize and search through all this information. Besides speech, our principal means of communication is through visual media and in particular through hand written documents. The automation of handwritten form processing is attracting intensive research interests due to its wide application and reduction of the tiresome manual workload. There are about 300 million people in India who speak and write Hindi Garhwali and Kumaoni, Konkani, Magahi, Maithili, Marwari, Bhili, Newari, and in other local languages. The documents have to be scanned and stored as images so that they may be processed by a computer. The textual content of these documents may also be extracted and recognized using OCR methods. An effective document processing system must be able to recognize structured and semi structured forms that is written by different people's handwriting. In this paper, we use an OCR method to read handwritten query by computer system and to provide required information using efficient Information Retrieval approaches.

Keywords- Information Retrieval OCR, Preprocessing, Artificial Neural networks, Segmentation, Feature Extraction, Classification

Citation: Pandey MK and Nandan Singh Dasila (2012) Information Retrieval Using Artificial Intelligence and Fuzzy Logic for Hand Written Documents through Optical Character Recognition (OCR). Journal of Information and Operations Management ISSN: 0976–7754 & E-ISSN: 0976–7762, Volume 3, Issue 1, pp-196-198.

Introduction

One may argue that computer literacy will eventually lead to paperless scenario in many large scale information processing systems, however such goal may not be achieved in near future because there are systems that largely include partially literate common people of rural India. Till today most of the Indian peoples are not competent in computer operation and there is various type of legacy requisition forms that needs to filled manually e.g. requisition for draft, reservation forms etc. The automation of handwritten form processing (from preformatted documents) is attracting intensive research interests due to its wide application and reduction of the tiresome manual Workload [1]. In Indian context most of the people, especially the senior people and the people from rural areas, are not yet familiar with personal computer/Internet system. Therefore online forms will not be able to completely replace physical forms (written and processed manually) in near future. Indian Railway Reservation System (IRRS) is an example of one such application [2]. It may be worth mentioning in this context that Indian railways are one of the largest in its category that caters around 14 million passengers a day. Although IRRS has an online reservation system but still most of the people fills up their railway reservation form manually. This work mainly focuses on handwritten query processing technology through OCR and its applications for information retrieval.

The value of computerized storage of handwritten documents would be greatly enhanced if they could be searched and retrieved in ways analogous to the methods used for text documents. If precise transcripts of handwritten documents exist, then information retrieval (IR) techniques can be applied; however, such transcripts are typically too costly to generate by hand, and machine recognition methods for automating the process of transcript generation can be implemented through OCR [3]. Optical Character Recognition is a process by which we convert printed document or scanned page to ASCII character that a computer can recognize. The document image itself can be either machine printed or handwritten, or the combination of two. Computer system equipped with such an OCR system can improve the speed of input operation and decrease some possible human errors. Difference in font and sizes makes recognition task difficult if preprocessing, feature extraction and recognition are not robust and width of the stroke is also a factor that affects recognition. Therefore, a good character recognition approach must eliminate the noise after reading binary image data, smooth the image for better recognition, extract features efficiently and classify patterns. In this paper, we propose a concept to read the handwritten forms in Devnagari script with the help of OCR system so that we can use it in Indian railways, Banking sectors and other rural areas of India for information retrieval.

Information retrieval deals with storage and processing of textual information. A basic task is here the retrieval of those documents from a collection which are relevant, i.e., match information needs of a user expressed as a query. The relevance may be meant as binary, i.e., a document is then regarded as either relevant or irrelevant. Thus the answer to a query is a set of documents considered to be relevant. More generally, a matching degree is computed for each document meant as an assessment of its relevance [4]. Then an answer to a query is a list of documents non-increasingly ordered against their matching degree. In our current work we have fill up the forms manually from random users. We then scan the filled-up forms using a flatbed scanner and analyze its contents using the custom-build form-recognition system. After scanning the hand written documents or forms, a fuzzy based IR system will be used to retrieve the information required by the user.

OCR System

Following steps have been followed in the OCR system:

- Preprocessing
- Segmentation
- Feature Extraction

Classification

Preprocessing

In the OCR system, text digitization is done by a flatbed scanner. The digitized images are usually in gray tone, and for a clear document, a simple histogram based threshold approach is sufficient for converting them to two tone images. The histogram of gray values of the pixels shows two prominent peaks, and a middle gray value located between the peaks is a good choice for threshold.

Segmentation

Segmentation is one of the most important phases of OCR system. Segmentation subdivides an image into its constituent regions or objects.

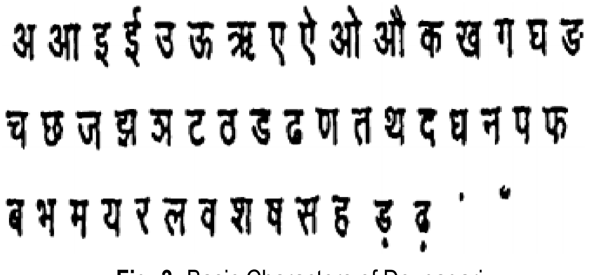

Fig. 3- Basic Characters of Devnagari

In segmentation, we try to extract basic constituent of the script, which are characters. This is needed because our classifier is able to recognize these characters only.

In Devnagari script, a text word may be partitioned into three zones as shown in figure 4. The upper zone denotes the portion above the headline, the middle zone covers the portion of basic and compound characters below the headline, and the lower zone may contain where some vowel and consonant modifiers can reside.

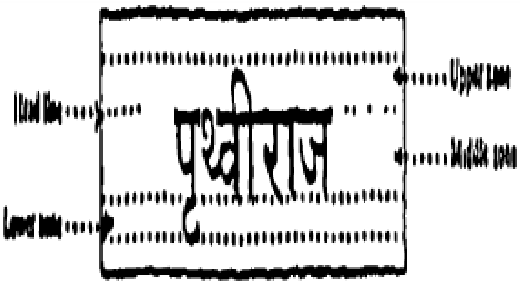

Fig 4: Partitioning of a Text Word into Zones

Classification

Classification is performed based on the extracted features. Here we are using ANN approach.

For initial classification of characters, we consider three features as follows:

- Mean Distance;
- Histogram of projection based on spatial position of pixel;
- Histogram of projection based on pixel value.

Artificial Neural Network (ANN) Approach for Classification

Artificial Neural Network approach has been used for classification and recognition. It is a computational model widely used in situation where the problem is complex and data is subject to statistical variation. Training and recognition phase of the ANN has been performed using conventional back propagation algorithm with two hidden layers. The architecture of a neural network determines how a neural network transfers its input into output. This transfer can be viewed as a computation

Feature Extraction

Feature extraction is one of the most important steps in developing a classification system. This step describes the various features selected by us for classification of the selected characters. Extraction of good features is the main key to correctly recognize an unknown character. A good feature set contains discriminating information, which can distinguish one object from other objects. It must also be as robust as possible in order to prevent generating different feature codes for the objects in the same class [5]. The selected set of features should be a small set whose values efficiently discriminate among patterns of different classes, but are similar for patterns within the same class. Features can be classified into two categories

Local features: Which are usually geometric (e.g. concave/convex parts, number of endpoints, branches, joints etc).

Global features: Which are usually topological (connectivity, projection profiles, number of holes, etc) or statistical (invariant moments etc.)

Query Processing Methodology

In our current work initially we scan the filled-up forms or hand written query of the user by using a flatbed scanner of the OCR system which uses Artificial Neural Network approach for classification and recognition. In the next stage, we try to retrieve the information from the given database using the fuzzy based information retrieval systems. The main steps involved in our system shall be as follows-

Reading the Hand written Query: OCR software is available for the Indian language Hindi (Devanagari), the third most spoken language in the world. Hindi characters based on the Devanāgarī script are distinguished by the presence of matras in addition to main characters. Matras are dependent vowels used for representing a vowel sound that is not inherent to the consonants. We have used the various steps of the OCR system like preprocessing, segmentation, feature extraction and classification. In preprocessing step it is expected to include noise removal, skew detection & correction. After finding out the feature of the segmented characters artificial neural network (ANN) [6], and [7] will be used for classification purpose. The use of artificial neural networks has improved the performance of the OCR systems. The OCR system shall be capable of accepting document images from a file or from a scanner directly and the recognized characters can also be displayed and edited.

Use of fuzzy logic for information retrieval: The basic task is the retrieval of those documents from a collection which are relevant, i.e., matching information needs of a user expressed as a query. We have used fuzzy logic based model, somehow inspired by all three traditional models, i.e. the traditional Boolean, vector space and probabilistic ones. We intend to obtain a comprehensive treatment of imprecision and uncertainty pervading the information retrieval process. Particular keywords represent the content of a document to a different degree. Usually, the notion of importance is used to describe the role of a keyword in this respect but it is not clear how to measure it. It may be conveniently assumed that importance is expressed by a real number from [0, 1]. To preserve the syntactical homogeneity of the representation of documents and queries, like in the classical Boolean model, a query is also represented as a compound linguistic statement [8]. Thus a linguistic assessment of selected keywords importance is given, each accompanied by a certainty degree. The statements are then transformed to a qualifier free form and the entire query is treated as a fuzzy set Q in a multidimensional space [0, 1]. The first step towards the application of fuzzy logic in IR is to employ multivalued logic instead of the binary one. First, a document is treated as a fuzzy set of keywords and the membership of a keyword reflects its importance in representing the meaning of the document. The next step is to allow for weights to be associated with keywords in a query [4]. This goes beyond the syntax of the classical, even multivalued, logic and calls for the use of an extended formalism in the context of the IR [9]. Basically, three most popular interpretations (semantics) of weights in queries are:

- **Relative importance**: If the weight of a keyword in a query is high, then its presence in a document (i.e., high weight there) is required for this document to match (to a high degree) the query.
- **Ideal weight**: The keyword is expected to have in a document a similar weight to that in the query.
- **Threshold:** The keyword is expected to have in a document a weight at least as high as that in the query.

To evaluate the relevance of a document against a query, the fuzzy model computes as in the classic Boolean model the truth degree of the statement q representing the query under the assumption that the statement d representing the document is true. According to Zadrozny's model the matching degree expresses the possibility and necessity of matching between a document and a query.

Conclusion

The current work is an attempt to simplify the existing manual data-processing for the hand written documents. The proposed query processing methodology will help in information retrieval in the field of scanned documents with effective use of fuzzy based systems. The future work will be implementing the use of dictionary words to improve the performance of OCR system. One can also implement the project for classifying hand-written text.

References

[1] Plamondon R. and Srihari S. (2000) *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 22(1):63 – 84.

[2] Rakshit S. and Das S. (2010) *International Journal of Computer Applications* (0975 – 8887)Volume 6– No.11.

[3] Russell G and Perrone M (2002) *International Workshop on Frontiers in Handwriting Recognition* © IEEE

[4] Zadeh L.A. (1975) *The concept of a linguistic variable and its application to approximate* reasoning. *Part I-III. Information Sciences,*8,8,9:199_249,301_357,43_80.

[5] Garain U and Chaudhary B. (2002) *IEEE Transaction on System, Man and Cybernetics- Part C:Applications and Reviews*, 32.

[6] Mori S et. al (1992) *IEEE*, 80, no 7, pp. 1029-1058.

[7] Chaudhary B. and Pal U. (1996) *IEEE 13th International Conference.* pp. 245-249.

[8] Zadrozny S., Nowacka K. and Kacprzyk J. (2008) *IPMU*, pp. 1749-1756.

[9] Zadrozny S and Kacprzyk J. (2005) *FUZZ-IEEE*, pp 1020-1025, Reno, NV, USA.

[10] Taghva K., Borsack J. and Condit A. (1994) *Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval,* pp. 202 – 211.

[11]Goyal A. and Khandelwal K. (2010) *Machine Learning*, *Fall'*10 CS229.