Suresh Gyan Vihar University,
Jaipur International Journal of Converging Technologies and Management (IJCTM)
Volume 1, Issue 2, 2015
ISSN : 2455-7528

A Classification Of Various Indian Script Recognition: A Survey

Divyang Vijayvergiya
Department Of Information and Technology
Gyan Vihar School of Engineering and Technology
Jaipur, India
divyangvijay1992@live.com

Akhilesh Pandey
Asst. Professor Department
Of Computer Science Engg.
Gyan Vihar School of Engineering and Technology
Jaipur, India
akhileshmtech10@gmail.com

Abstract

Till now, many researches have been done for optical character recognition (OCR) and a number of articles have been published in the last few decades. In market a number of commercial OCR systems are available. But these OCR systems are developed for Chinese, Arabic, Japanese and roman but only few are available for Indian languages though there are 12 major scripts in India. This paper provides a review for the OCR work done on Indian languages scripts. This review is arranged in some sections. Section 1 contains Introduction part. In section 2, Indian script Features are discussed and methodologies in OCR system is discussed in section 3. Section 4 concludes the paper.

1. Introduction

Optical Character Recognition has been one of the most demanding research area in the field of pattern recognition in the recent years. A number of researches has been done in Pattern recognition to develop newer techniques which would decrease the processing time at the same time providing higher recognition accuracy. OCR is a technique by which characters are automatically recognized by computer in optically scanned and digitized pages of text. It helps in improving interface among human and machine in a number of applications like automatic text entry into computer for desktop publication, document data compression, automatic reading of bank The idea of character recognition was first developed in 1870. At that time Carey invented the retina scanner – an image transmission system using a mosaic of photocells [1]. After two decades, sequential scanner was developed by nipkow which provided a major advancement in reading machines and modern televisions. However, optical Character recognition was primarily considered to support visually handicapped persons and first successful attempt was made by the Russian scientist Tyurin in 1900.

OCR systems can be separated into four generations. The First generation systems involve constrained letter shapes that the OCR system reads. Such machine marked their presence in the starting of 1960s. IBM 1418 was the first commercialized OCR of this generation [2]. Logical template matching was the recognition method working behind these machines where the positional relationship was fully utilized.

The Second Generation was capable of recognizing a set of machine printed characters along with hand written characters. Such Systems were considered to be hybrid because of the combination of analog and digital technology.

The third generation can be marked by the OCR which involve poor print quality and hand- printed characters

cheques and other documents etc. Optical Character ex: fax, photocopy etc, color document, unconstrained hand
Recognition process can be sub-divided into following five written characters. stages:

At the present time, more sophisticated OCRs are available

1. Pre-processing for Chinese, Japanese, Arabic and roman text. Documents that
2. Segmentation has been typewritten or printed by dot matrix, laser and line
3. Feature Extraction printers. Characters with different sizes and font are recognize
4. Classification using these OCRs.
5. Post-processing

`

Suresh Gyan Vihar University,
Jaipur International Journal of Converging Technologies and Management (IJCTM)
Volume 1, Issue 2, 2015
ISSN : 2455-7528

The fourth generation is marked by the OCR which involve intermixing of complex document with graphics, table, text and mathematical symbols, low quality noise documents for

2. Indian scripts features

Among all the languages of india, Hindi is the most popular whereas Bangla is at second position in india and are third and fourth most popular language in the world. For writing Indian official (Indian constitution accepted) language, twelve different scripts are used. By applying a variety of alteration on ancient Brahmi many of Indian scripts are formed [7]. Devnagari is most popular script in india as it is used to write many different languages like Hindi, Sanskrit, Rajasthani, and Nepali, whereas Bangla script is used to write Bengali and Assamese languages.

Consonant and Vowel characters are called basic characters, whereas most Indian script also consist of compound characters(formed from combination of two or more basic characters) as well. Compound characters are generally more complex than constituent basic characters. A vowel can be located to the left, right, top, bottom of the consonant in some languages. They are known as modified characters. In Indian script there are nearby 300 character shapes [8].

Many characters in some Indian script alphabets (for ex: Devnagari, Gurumukhi, etc.) have horizontal line at higher part. This line is known as sirorekha in Devnagari and matra in Bangla. To form a word, two or more characters sits together, the head line portion of word touches one character to another form a big head line. Due to this head-line, character segmentation for Optical Character Recognition is required. Some scripts like Oriya, Gujarati etc do not have head-line.

A text line can be partitioned into three portions in most of Indian languages. Portion above the head line can be said as upper portion, the middle portion below the head-line consist of basic and compound characters and then the final lower portion.

Unlike English, Indian languages does not have upper and lower case characters. But writing mode is similar as English i.e, from left to right. However, Urdu script is written from right to left. Also Urdu script comes under Perso Arabic group of scripts.

3. OCR methods and work done on Indian scripts

Earlier, pattern recognition techniques are classified as feature and template based approach. An unknown pattern is directly superimposed on the suitable template pattern and similarity between the two identifies the classification. Earlier, only template based approach was used in OCR systems, but recent systems uses combination of both template and feature based approach which provides better results. For example, feature based approach is used in Bangla OCR system [8] for basic characters recognition and template based approach is used for compound character recognition.

Feature based approach extract important features from the test patterns and employ them into a extra sophisticated model of classification. The feature based approach is further divided into two types, namely transform domain and spatial domain approaches. In spatial domain approaches features are directly derived from the representation of pixel in pattern whereas in transform domain technique, pattern image is transformed to another space by using Cousine, Slant, Wavelet or Fourier transformation and then important features are extracted from transformed images. Spatial domain features OCR are mainly used for different scripts like Devnagari, Telugu, Bangla etc. Moment based [9,10], Syntactic grammar [9], and graph theoretic approaches are also tested for the OCR problems.

Modern techniques do not explicitly extract any feature directly from the patterns [3,12]. Systems are fed with normalized or raw pattern during training which improves systems at its misclassification error of patterns. A good example of this type of system is artificial neural network which adjust its links weight from training pattern.

Not much work is done on Indian language script recognition. Most of the existing work concerns about Bangla and Devnagari script. Some studies have also been done on recognition of other Indian languages like Gujarati, Telugu, Tamil, Punjabi, etc. Neural network classifier, structural and topological features based tree classifier are mostly used for Indian script recognition. Script wise work on Indian languages are reviewed below.

`

Suresh Gyan Vihar University,
Jaipur International Journal of Converging Technologies and Management (IJCTM)
Volume 1, Issue 2, 2015
ISSN : 2455-7528

3.1 Work done on Devnagari script recognition

For optical character recognition of Devnagari script, Sinha and Mahabala [11] offered a syntactic pattern analysis system which consist of embedded picture language for recognition of handwritten and machine written characters of Devnagari. Structure of each symbol of Devnagari script is stored in system in terms of primitives and their relationships. For recognition, input character is marked and is compared with stored description. Related information about the occurrences of certain primitives, and their combination is used to improve the accuracy of system and reduce the cost of computation.

Fig: An example of Devanagari script for Hindi Language

The first complete development of OCR system for printed Devnagari is possible because of Pal and Chaudhuri. Some standard techniques were used and some new methods have been developed for development of this OCR system by them. In Devnagari, a long line is generated when we form a word using two or more characters, is called a head-line. Due to this head-line, segmentation of characters become difficult. In this approach, deletion of head-line is made for segmentation. And text line is separated into three horizontal zones. Basic, compound and modified characters are separated from the zonal information and shape properties. Structural feature based tree classifier is used to recognize basic and modified characters whereas hybrid approach which is combined with structural and run based template feature is used to recognize compound characters. It gives around 96% accuracy.

3.2 Work done on Bangla script recognition

The First complete OCR system for Bangla documents was developed by Chaudhuri and Pal [8]. In this, in preprocessing, skew correction is carried out followed by removal of noise, and segmentation of input image into line, characters and zones. A combination of template and feature matching is employed for recognition of Bangla script. There are mainly eight stroke based features and for dot representation, a filled circle. Feature based tree classifier recognizes simple characters and compound characters are recognized by run based template matching preceded by feature based grouping. Individual character occurrence frequency, bigram and trigram statistics have been utilized to support the recognition process. 96% accurarcy is provided by this system.

Garain and Chaudhuri [17] proposed a technique for recognition of printed Bangla characters which combines the positive aspects of feature and run number based normalized template matching technique. For horizontal and vertical scanning, run number vectors are computed. For different patterns, number of scan also vary, they normalized and abbreviated the vector. They proved that normalized and abbreviated vector induces a metric distance and observed that this method is more helpful for complex shaped characters than simple ones. In a group of compound characters, for matching, they apply vector representation. They observed that vector is more efficient if vector is restructured with regard to centroid of the pattern.

3.3 Work done on Telugu script recognition

Rajasekaran and Deekshatulu [16] proposed a two level recognition system for printed Telugu characters. First level uses a knowledge based search to identify and eliminate primitive shapes present in the input alphabet. For this purpose, a directed curve tracing technique is applied. Then the pattern achieved after the elimination of primitives is coded using tracing beside the points on it, in the second level. Decision tree provides the classification on the basis of knowledge obtained about primitives and basic characters in the input.

3.4 Work done on Tamil script recognition

For Tamil character recognition of hand printed documents, Chinnuswamy and Krishnamoorthy [18] presented an approach in which alphabets are assumed to be made of line like elements, known as primitives and satisfies some relational constraints. The structural composition of characters are represented in terms of primitives using labeled graphs. In recognition process, conversion of input image into labeled graph is carried out which represents the input character and for a group of basic symbols, labeled graph with computing correlation coefficients is stored. Topological matching

technique is use to compute the correlation coefficients and then correlation coefficient is maximized.

3.5 Work done on Gurumukhi script recognition

Gurumukhi script is similar to Devnagari but is simpler because compound characters are absent in the script. For printed Gurumukhi script a complete OCR system has been developed by Lehal and Singh [14], in which segmentation is first applied on connected components using thinning based approach. Two types of features are there in this process. First is primary feature which sets the number of loops, number of junction and their positions are tested. Second is secondary feature which consist of nature of profiles of different directions, number of end point and their location etc. For this purpose a multi stage classification scheme combined with nearest neighbor and binary tree classifier is used. It has 97.34% accuracy.

3.6 Work done on Gujarati script recognition

Ananti and Agnihotri [13] proposed a classification for a subset of printed Gujarati alphabets. For classification, K nearest neighbor classifier was used with usual and invariant moments. Hamming distance classifier has been also used. The recognition rate was observed nearby 67%.

3.7 Work done on Oriya script recognition

In Oriya script, a large number of characters are present in which many characters are of similar shapes, that is why to build OCR system for Oriya script is difficult. Only a few work have been done on recognition of Oriya alphabets.

Mohanti [15] presented a system using kohonen neural network to recognize characters of Oriya script. Input pixels are inserted to the neurons present in Kohonen layer where using weighted sum formula, neurons identifies the output. The alphabet is recognized by the maximum output obtained from the neuron. In this, author made experiment on only five characters of Oriya, therefore the system is not reliable.

4. Conclusion and future Scope

In this paper a review of OCR work done on different Indian language scripts is presented. Different methodologies used in OCR development and work done on different Indian language script recognition is described. We believe that this survey will encourage the activities of automatic processing of documents and OCR of Indian language scripts.

References

[1] J. Mantas, An overview of character recognition Recogn. Artif. Intell.12 (1998) 1007-1015.

[5] L. O' Gorman, R. Kasturi, Document Image Analysis, IEEE Computer Society Press, Los Alamitos, CA, 1995.

[6] W. Stallings, Approaches to Chinese character recognition, Pattern Recognition 8 (1976) 87-98.

[7] A.K. Dutta, A generalized formal approach for description and analysis for major Indian scripts, J. Inst. Electronic Telecom. Eng. 30 (1984) 155-161.

[8] B.B. Chaudhuri, U.Pal, A complete printed Bangla OCR system, Pattern Recognition 31 (1998) 531-549.

[9] F. Feng, T. Pavlidis, Decomposition of polygons into simpler components: feature generation for syntactic pattern recognition, IEEE Trans . Comput. 24 (1975) 636-650.

[10] S.S. El-Dabi, R. Ramis, A. Kamel, Arabic character recognition system: a statistical approach for recognizing cursive type-written text, Pattern Recognition 23 (1990) 485495.

[11] R.M.K. Sinha, H. Mahabala, Machine recognition of Devnagari script, IEEE Trans Systems Man Cybern. 9 (1979) 435-441.

[12] C. Choisy, A. Belaid, Cross-learning in analysis in analytic word recognition without segmentation, Int. J. Document Anal. Recognition 4 (2002) 281-289.

[13] S. Antani, L. Agnihotri, Gujrathi character recognition, Proceedigns of Fifth International Confrence on Document Analysis and Recognition, 1999, pp. 418-421.

[14] G.S. Lehal, C. Singh, Feature extraction and classification for OCR of Gurumukhi script, Vivek 12 (1992) 2-12.

[15] S. Mohanti, Pattern recognition in alphabets of Oriya language using Kohonen neural network, Int. J. Pattern

`

methodologies, Pattern recognition 19 (19860) 425-430.

[2] S. Mori, C.Y. Suen, K. Yamamoto, Historical review of OCR research and development, Proc. IEEE 80 (1992) 10291058.

[3] R. Plamondon, S.N. Srihari, On-line and off-line handwritten recognition: a comprehensive survey, IEEE Trans. Pattern Anal. Mach. Intell, 22 (2000) 62-84.

[4] A. Amin, Off-line Arabic character-recognition: the state
of the art, Pattern Recognition 31 (1998) 517-530.

[16] S.N.S. Rajasekran, B.L. Deekshatulu, Recogniton of printed Telugu characters, Comput. Graphics Image Process. 6 (1977) 335-360.

[17] U. Garain, B.B. Chaudhuri, Compound character recognition by run-number-based metric distances, SPIE Proc. 3305 (1996) 90-97.

[18] P. Chinnuswamy, S.G. Krishnamoorty, Recognition of hand-printed Tamil characters, Pattern Recognition 12 (1980) 141-152.

`