International Journal of Emerging Research in Management &Technology
ISSN: 2278-9359 (Volume-6, Issue-6)

Research Article | **June 2017**

A Study of Different Methodologies Helpful in the Identification of Offline Handwritten Script

Manish M. Kayasth*
Asst. Prof. and HOD, UCCC & SPBCBA &
SDHG College of BCA and IT,
Surat, Gujarat, India

Bharat C. Patel
Asst. Prof. and I/c Principal, T. & M. T.
College of Information Science,
Surat, Gujarat, India

Abstract –

The entire character recognition system is logically characterized into different sections like Scanning, Pre-processing, Classification, Processing, and Post-processing. In the targeted system, the scanned image is first passed through pre-processing modules then feature extraction, classification in order to achieve a high recognition rate. This paper describes mainly on Feature extraction and Classification technique. These are the methodologies which play an important role to identify offline handwritten characters specifically in Gujarati language. Feature extraction provides methods with the help of which characters can identify uniquely and with high degree of accuracy. Feature extraction helps to find the shape contained in the pattern. Several techniques are available for feature extraction and classification, however the selection of an appropriate technique based on its input decides the degree of accuracy of recognition.

Keywords - Feature Extraction, Classification, Gujarati Script

I. INTRODUCTION

The current generation of Optical Character Recog-nition (OCR) systems can be characterized as a pipeline composed of Preprocessing, Segmentation, Classification, and Identification stages. None of these stages are immune to error. Preprocessing may fail to remove existing noise, it may remove portions of the image or add noise by some other mechanism. Segmentation may fail to establish a boundary where there should be one (joining error), it may mistakenly introduce a boundary where there should not be one (splitting error), it may ascribe the wrong co-ordinates to a boundary (misalignment error), or display any combination of these errors over multi-segment stretches of text. Classification may be mistaken (substitution error) or may provide no output at all (rejection error). Identification of significant units (words, phrases, etc.) may fail because of low quality character-level input or because of inadequacies in the system dictionary or context model [12].

Because the development of reliable classifiers requires considerable engineering effort, and is still not an entirely solved problem, most commercial efforts in OCR concentrate on machine print, forms with pre-set character boxes, or discrete handwriting styles where the segmentation problem is less acute [12]. Handwriting recognition has been a subject of research for several decades [13].

The research on identifying characters in handwriting has made progress and many such systems have been developed in that area especially in online character recognition and machine printed typed characters. However, it is not true in the recognition of handwritten characters, especially for offline. The character recognition rate is dependent on writing habits and the machine used in the recognition of offline handwritten text. In the recent years there have been attempts made to deal with uncertain and incomplete information due to confusing shapes and writer variability using Hidden Markov Models. In fact it is difficult to process handwritten characters because of great variations in writing, size of text, writing angles of the characters etc.

Generally, a character recognition system includes three main tasks: preprocessing, feature extraction, and classification. In pre-processing, researchers normally perform noise filtering, binarization, thinning, skew correction, slant normalization, etc. to enhance the quality of images and to correct distortion. It is observed that low resolution of original data could reduce the recognition rates of OCR systems dramatically [13]. Matrix matching works best when the OCR encounters a limited collection of type styles, with little or no variation within each style. In general, the matching of numbers to a template or pattern may be too time-consuming and not flexible enough therefore feature extraction is needed [1]. It is also found that recognition rate is drastically reduced without classification [10]. However, the development of reliable classifiers requires considerable effort, and is still not an entirely solved problem [12].

The feature extraction phase analyses the given text segment and selects a set of features that can be used to uniquely identify the given text segment. In the classification section of this phase, features extracted in the feature extraction section are used to identify the text on the basis of pre-decided rules. The output of the OCR system generally contains errors; the debugging responsibility is performed by the last phase that is Post processing phase. Thus all phases of the OCR system are very important, but the most important phase which is responsible for character recognition is Feature Extraction and Classification [6].

Among various existing languages it is being easier to recognize English text and numerals, however it is not so for the few Indian languages like Gujarati. An intensively for the recognition of English, Chinese, Japanese languages and hand written numerals though less attention has been given to Indian language recognition. Some efforts have been reported in literature for Devnagari Tamil scripts, Hindi numerals and printed Telugu texts. Some work has been reported in recognition of unconstrained handwritten characters using different number of classes and different classification strategies. Neural network is a traditional recognition technique used.

Several algorithms are available for the character recognition. However, most of their work is related to English text. The areas of Handwriting recognition have seen very little research in the context of Indian scripts. The inflectional and agglutinative nature of Indian languages makes the OCR task quite challenging. Preliminary results are available in the literature on the recognition of two popular scripts of south India – Tamil and Kannada. Very few works have been reported in the language like Gujarati [2], [3].

Gujarati is a language from the Indo-Aryan family of languages, used by around 50 million people in the western part of India. It is one of the India's most popular languages mainly used in the Gujarat state. Gujarati-script used to write the Gujarati language. Gujarati Script has a large basic set of characters. It utilizes 75 distinct legitimate and recognized shapes and 45 conjuncts and special characters. It mainly includes 59 characters and 16 diacritics. 59 characters are divided into 36 basic consonants (34 singular and 2 compound (not lexically though)) means ornamented sounds, 13 vowels (pure sounds), and 10 numerical digits, whereas 16 diacritics are divided into 13 vowel and 3 other characters. The alphabet is ordered by logically grouping the vowels and the consonants based on their pronunciations. It has half and special characters which are used with consonants and vowels to form and enhance the word vocabulary. The Gujarati text can be partitioned into basic and compound characters surrounds the core character. The Gujarati script assigns unique UNICODE numbers by Web Consortium i.e. U+0A80 – U+0AFF (02688 – 02815) As per the best of authors' knowledge, there is no ready and reliable work available on recognition of Handwritten Gujarati language [2], [3].

The paper narrates Image Enhancement and Thresholding in section II, Feature extraction technique in Section III, and Classification in section IV. Finally, we conclude and analyze the effect of feature extraction and classification on the recognition of handwritten Gujarati characters with future attempts.

II. IMAGE ENHANCEMENT AND THRESHOLDING

In image enhancement, the goal is to accentuate certain image features for subsequent analysis or for image display. Examples include contrast and edge enhancement, pseudo coloring, noise filtering, sharpening and magnifying. Image enhancement is useful in feature extraction, image analysis and visual information display [4].

The task of thresholding is to extract the foreground from the background. There are two techniques proposed for thresholding: global and local. Global methods apply one threshold to the entire image while local thresholding apply different threshold values to different regions of the image.

III. FEATURE EXTRACTION

After preprocessing stage an image has been segmented into regions, the image is subjected to feature extraction where different feature extraction methods are available. In order to decide the identity of the text, features, which will exhibit the distinct characteristics of those particular characters, are extracted. Ideally, the features should enable the system further, to discriminate or separate one class of text to other classes correctly. In addition, the central issue in the recognition is the selection of proper features [10]. However, the selection of features for character recognition can be problematic by the fact that piece of document or text may have different sizes, shapes, layouts etc. makes process more complicated.

Feature extraction is a method used for OCR without strict matching to prescribed templates. It is also known as topological feature analysis. This method varies by how much computer intelligence is applied by its creator. This method is much more versatile than simple matrix matching. Feature or topographical analysis is superior when the characters are less predictable [14].

The process of handwriting recognition involves the extraction of some define characteristics called features to identify unknown characters. Feature extraction method plays an important role in the success of hand written character recognition system. Here the features are extracted from a closed boundary trace of a character. There are many features that can be used to describe a closed boundary trace.

In character recognition, it is desirable to extract features or say raw data from the text which are focused on discriminating between classes. Feature extraction can be considered as finding a set of vectors, which effectively represent the information content of a character [8]. More specifically, it looks for general features such as open areas, closed shapes, diagonal lines, line intersections, etc. in the character image [14]. It is a technique for extracting such raw data which is most relevant or significant for classifying a character. In other words, it is used specifically to avoid character misclassification and in this way it increases high recognition rate [8]. The following is the Gujarati numeral '5' which can be misclassified easily as Gujarati character 'p' as shown in the following figure 1.

Number 5 ('5') Character p ('p')

Fig 1: Number '5' and character 'p' in Gujarati

Various types of features extraction techniques are available, such as geometric features, wavelet features, moments, quantities derived from series expansion, structural features, component analysis etc. [9], [13]. These methods are analyzed based on Nature of input image, and the type of classifier used. Which feature extraction method is the best to achieve high recognition for a given application is an important issue [8]. Different feature extraction methods are designed for different representations of the characters, such as solid binary characters, character contours, skeletons or thinned characters, gray-level sub-images of each individual character etc. [5].

Feature extraction technique obtains structural features i.e. shape of each character fragment from the image [1]. The methods implemented different stages like: (1) Zoning (2) Center of gravity (3) Tips (4) Number of branches or intersections (5) Projection histograms (6) Extreme points (7) Direction of contour pixels (8) Wavelets (9) Fourier descriptors and (10) Moments [8].

The authors propose the following strategy for the feature extraction task: The Feature Extraction task involves the extraction of meaningful characteristics of a segmented character. First it divides the character image into several frames using a certain frame length. Then it performs Discrete Cosine Transform (DCT) calculation over each frame. Finally, consider the number of frames and DCT calculated values of each frame as the features for each character [11]. A recent feature extraction method, namely sector data method takes care of variability involved in the writing styles of individuals. This sector based approach is better suited for the recognition of Gujarati Numerals when classification is made on the basis of strokes [10].

IV. CLASSIFICATION

After feature extraction method described in section III, the feature vector is fed into a classifier. Each character is assigned to one of the given classes. As the recognition of characters is a daunting task, raw or common classification is necessitated. The classification is carried out at the final stage to recognize the character. It assigns an input character to one of many pre-specified classes which are based on the extracted features and their analysis [1]. In addition, it is next to impossible to get correct classification with only one classifier in view of the fact that some of the characters have similar shapes.

A number of classification methods are available. Some of them are feature based tree classifier, nearest neighbor classifier [7], Artificial Neural Networks (ANN), Support Vector Machines (SVM), Multiple Classifier Systems, statistical classifiers, Syntactic or Structural methods, Template matching etc. [6], [13]. ANN and SVM are classifiers use for analysis because they have good learning ability and exhibited good performance. Other two different classifiers like a 3-layer ANN with back propagation algorithm and SVM classifier are used to test all the patterns [13].

There are different natures of classifiers used like: (1) Structural classifier which tests whether particular features exist in specific positions within character region means geometric positions or shapes. E.g. Detecting edges (using Canny's algorithm), corners, straight lines etc. (2) Statistical classifier which uses statistical techniques using training set means Real-valued features. E.g. Baye's decision making, nearest neighbor classifier (3) Decision trees uses distinct features (4) Multistage classifiers imply same or different features at each stage [8]. Ahmadi etal. and Cho have employed a multistage classifier for recognizing unconstrained handwritten numerals [10].

The authors suggest the Neural Networks based classifier where result is good for the segmented characters. However any such classifier is highly sensitive to the quality of the character images given as input. Therefore it is essential that the preprocessing components of the system like character image extraction, segmentation etc. are well designed.

The system is integrating features of characters such as aspect ratios, pixel density, number and position of points. It also uses structural descriptions or features of character and the holistic features in the classifier. The structural features of characters uses are the location of vertical and/or horizontal bar, connectivity of character components, and which side the characters are open to, close to etc. Further, the holistic features like word length, number and positions of ascenders, descenders and loops are used. These techniques are used to classify characters and so followed by performing the recognition.

V. CONCLUSION AND FUTURE ATTEMPTS

In this paper, authors have presented processes of feature extraction and classification techniques for OCR of handwritten scripts. A lot of research has been done in this field. Still the work is going on to improve the accuracy of feature extraction and classification techniques. The different methods of feature extraction and classification described in the paper are very effective and useful for further research in the same area. The methods presented here can be further extended by post processing activities for the recognition of Gujarati handwritten characters.

REFERENCES

[1] Mohsen Zand, Ahmadreza Naghsh Nilchi et al., Recognition-based Segmentation in Persian Character Recognition, International Journal of Computer and Information Science and Engineering 2, pp. 14-18, 2008.

[2] Manish M. Kayasth, Bankim Patel, Offline Pre-Processing Operations for Recognition Of Handwritten Characters In Gujarati Language, International Journal of Computer Applications in Engineering Technology and Sciences, vol. 4, No. 1, pp. 102-106, Oct. 2011.

[3] Manish M. Kayasth, Novel Pre-Processing Operations for Detection of Handwritten Gujarati Characters, International Journal of Computer Science and Technology, vol. 5, Issue 1, pp. 9-13, Jan.- Mar. 2014.

[4] Ambika Ramchandra, Filters a Image Enhancement and Smoothing Technique, Paripex – Indian Journal Of Research, pp. 31-33, vol. 2, Issue:7, 2013.
[5] Trier Due, Anil K. Jain et al., Feature extraction methods for character recognition, A survey, Pattern Recognition, vol. 29, issue 4, pp. 641-662, February, 1999.
[6] Rohit Verma, Dr. Jahid Ali, A-Survey of Feature Extraction and Classification Techniques in OCR Systems, International Journal of Computer Applications & Information Technology vol. 1, Issue 3, pp. 1-3, November 2012.
[7] C. Weliwitage, A. Harvey et al., Whole of Word Recognition Methods for Cursive Script, WDIC, pp. 111–116, February 2003.
[8] Rathna Ramaswamy, Performance Evaluation Of Feature Extraction Algorithms For Character Recognition, April, 2006.
[9] Hriday Ravindranath, Written Letter Analysis, 2008.
[10] M. Hanmandlu, M. Vamsi Krishna et al., Some Approaches To The Recognition Of Handwritten Numerals.
[11] N. Papamarkos, J. Tzortzakis et al., Determination of Run-Length smoothing values for document segmentation, 3rd IEEE International Conference, vol. 2, pp. 684-687, 1996.
[12] Andras Kornai, K.M. Mohiuddin et al., An HMM-Based Legal Amount Field OCR System for Checks, In Proceedings IEEE International Conference on Systems, Man and Cybernetics, vol. 3, pp. 2800- 2805, 1995.
[13] Chun Lei He, Ping Zhang et al., The Role of Size Normalization on the Recognition Rate of Handwritten Numerals, Proceedings Neural Networks and Learning in Document Analysis and Recognition, August, 2005.
[14] http://www.dataid.com/aboutocr.htm

