Gujarati Character Recognition: A Brief Survey

Sujata Anandwani[1] Nirali Makad[2] Nidhi Gondalia[3]
[1]P.G. Student [2,3]Assistant Professor
[1,2,3]Departement of Computer Engineering
[1,2,3]Noble Group of Institution, Gujarat, India

Abstract— English Character Recognition techniques expanded in the field of pattern recognition research in the recent years and its progress is quite high. But for Indian regional languages are it is still emerging approach and progress is too slow. In Gujarat, so many of people use Gujarati in speaking and writing. Nowadays the entire world is digitized. We can see substantial demand of digital documentation in any field like postal services, data entry, publishing house, automation, and banking systems. In the field of research neural network is one of the largest domain that can be implemented in number of day to day activity problems. One of the important application of neural network is classification. Classification means shorting of any data or entity. It may be using computer, classification of peanuts, classification of letters in post office, recognition of disease from MRI image, or it may be classification or recognition of characters. A brief history of OCR, various approaches and progress to character recognition for Gujarati Language along with their status is also discussed in this paper.

Key words: Pattern Recognition, Optical Character Recognition, Template matching, Artificial Neural Networks

I. INTRODUCTION

Today technology is spanned in very aggressive manner all over the world. Cheap price of computer and internet has given acceleration to reach every corner of the world. Today most of documents systems in government offices or everywhere is text based. Large amount of documents and literature are in printed text format or in scanned format. There is need of some efficient method that can identify characters from the printed scanned documents. A computer system that recognizes characters from a scanned image or document and can process automatically is called Optical Character Recognition (OCR) system. One of the initial techniques is Template Matching technique.

In English it all about 26 alphabets and lots of work is done for English handwritten and printed character recognition but if we talk about Indian scripts, which contain partial characters, joint characters and lots of similar characters, at that time all these developed methods for English are not useful for Indian scripts. Brahmi scripts are far more complex than English scripts. All the Indian languages lie under Brahmi script.

II. FEATURES OF GUJARATI LANGUAGE

Gujarati is an official language of western part of India. It has 34 consonants which are also known as 'Vyanjans' figure (1) and 13 vowels figure (2) called 'Swar' as below. [22]

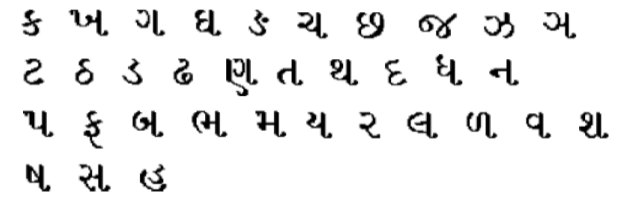

Fig. 1: Gujarati Consonants

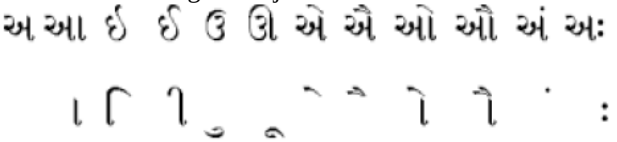

Fig. 2: Gujarati Vowels and modifiers

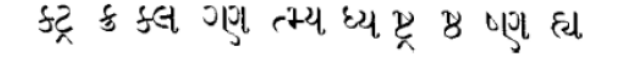

Fig. 3: Gujarati Conjuncts

III. OPTICAL CHARACTER RECOGNITION

The main principle in automatic recognition of patterns is first to teach the machine which classes of patterns that may occur and what they look like. In OCR the patterns are letters, numbers and some special symbols like commas, question marks etc., while the different classes correspond to the different characters. The teaching of the machine is performed by showing the machine examples of characters of all the different classes. Based on these examples the machine builds a prototype or a description of each class of characters. Then, during recognition, the unknown characters are compared to the previously obtained descriptions, and assigned the class that gives the best match.

In most commercial systems for character recognition, the training process has been performed in advance. Some systems do however; include facilities for training in the case of inclusion of new classes of characters.

A. Components of an OCR system:

A typical OCR system consists of several components. In figure 4 a common setup is illustrated.

The first step in the process is to digitize the analog document using an optical scanner. When the regions containing text are located, each symbol is extracted through a segmentation process. The extracted symbols may then be preprocessed, eliminating noise, to facilitate the extraction of features in the next step.

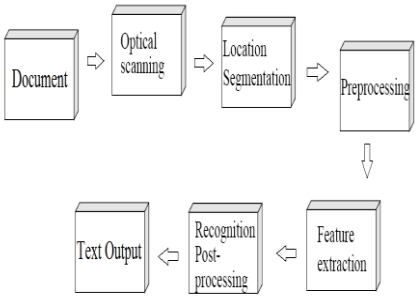

Fig. 4: OCR system

The identity of each symbol is found by comparing the extracted features with descriptions of the symbol classes obtained through a previous learning phase. Finally contextual information is used to reconstruct the words and numbers of the original text. In the next sections these steps and some of the methods involved are described in more detail. [21]

B. Optical Scanning:

Through the scanning process a digital image of the original document is captured.

In OCR optical scanners are used, which generally consist of a transport mechanism plus a sensing device that converts light intensity into gray-levels. Printed documents usually consist of black print on a white background. Hence, when performing OCR, it is common practice to convert the multilevel image into a bi-level image of black and white. Often this process, known as thresholding, is performed on the scanner to save memory space and computational effort.

C. Location And Segmentation:

Segmentation is a process that determines the constituents of an image. It is necessary to locate the regions of the document where data have been printed and distinguish them from figures and graphics. For instance, when performing automatic mail-sorting, the address must be located and separated from other print on the envelope like stamps and company logos, prior to recognition.

Applied to text, segmentation is the isolation of characters or words. The majority of optical character recognition algorithms segment the words into isolated characters which are recognized individually. Usually this segmentation is performed by isolating each connected component that is each connected black area. This technique is easy to implement, but problems occur if characters touch or if characters are fragmented and consist of several parts. The main problems in segmentation may be divided into four groups:

- Extraction of touching and fragmented characters.
- Distinguishing noise from text.
- Mistaking graphics or geometry for text.
- Mistaking text for graphics or geometry.

D. Preprocessing:

The image resulting from the scanning process may contain a certain amount of noise. Depending on the resolution on the scanner and the success of the applied technique for thresholding, the characters may be smeared or broken. Some of these defects, which may later cause poor recognition rates, can be eliminated by using a preprocessor to smooth the digitized characters.

The smoothing implies both filling and thinning. Filling eliminates small breaks, gaps and holes in the digitized characters, while thinning reduces the width of the line. The most common techniques for smoothing move a window across the binary image of the character, applying certain rules to the contents of the window.

In addition to smoothing, preprocessing usually includes normalization. The normalization is applied to obtain characters of uniform size, slant and rotation. To be able to correct for rotation, the angle of rotation must be found. For rotated pages and lines of text, variants of Hough transform are commonly used for detecting skew. However, to find the rotation angle of a single symbol is not possible until after the symbol has been recognized. [21]

E. Feature extraction:

The objective of feature extraction is to capture the essential characteristics of the symbols, and it is generally accepted that this is one of the most difficult problems of pattern recognition. The most straight forward way of describing a character is by the actual raster image. Another approach is to extract certain features that still characterize the symbols, but leaves out the unimportant attributes. The techniques for extraction of such features are often divided into three main groups, where the features are found from:

- The distribution of points.
- Transformations and series expansions.
- Structural analysis.

F. Classification:

The classification is the process of identifying each character and assigning to it the correct character class. For this two different approaches for classification in character recognition are used. First decision-theoretic recognition is treated. These methods are used when the description of the character can be numerically represented in a feature vector.

It may also have pattern characteristics derived from the physical structure of the character which are not as easily quantified. In these cases the relationship between the characteristics may be of importance when deciding on class membership. For instance, if we know that a character consists of one vertical and one horizontal stroke, it may be either an “L” or a “T”, and the relationship between the two strokes is needed to distinguish the characters. A structural approach is then needed.

G. Post Processing:

1) Grouping:

The result of plain symbol recognition on a document is a set of individual symbols. However, these symbols in themselves do usually not contain enough information. Instead we would like to associate the individual symbols that belong to the same string with each other, making up words and numbers. The process of performing this association of symbols into strings, is commonly referred to as grouping. The grouping of the symbols into strings is based on the symbols’ location in the document. Symbols that are found to be sufficiently close are grouped together.

2) Error-detection and correction:

Up until the grouping each character has been treated separately, and the context in which each character appears has usually not been exploited. However, in advanced optical textrecognition problems, a system consisting only

of single-character recognition will not be sufficient. Even the best recognition systems will not give 100% percent correct identification of all characters, but some of these errors may be detected or even corrected by the use of context.

There are two main approaches, where the first utilizes the possibility of sequences of characters appearing together. This may be done by the use of rules defining the syntax of the word, by saying for instance that after a period there should usually be a capital letter.

Another approach is the use of dictionaries, which has proven to be the most efficient method for error detection and correction. Given a word, in which an error may be present, the word is looked up in the dictionary. If the word is not in the dictionary, an error has been detected, and may be corrected by changing the word into the most similar word. [21]

IV. Literature Survey

A. Earlier Attempt For Gujarati Character Recognition:

This approach describes the results obtained by using the Euclidean Minimum Distance classifier, the *k*–Nearest Neighbor classifier, and the Hamming Distance classifier. The features used for these classifiers are the regular moments, invariant moments and the distribution in the binary feature space.

This system does not have the usual preprocessing phases that separate words from sentences and characters from words. It also does not have skew correction or noise removal etc. These are preprocessing phases in a typical text reading system. This was not accurate in case of various resolution and font family. [18]

Fringe distance is used as distance measure for the comparison of Gujarati character binary images. It is assumed that the characters are in black on a white background. Fringe distances compare only black pixels and their positions between the templates and the input images. Fringe distances may be even more efficiently computed by precompiling and storing the distances of the nearest black pixel at each pixel position of the template. This is called the fringe distance map. When input is compared to a template, the fringe distance map of the input character is computed and superimposed upon the template. A character, with the minimum fringe distance, is said to be recognized by the template. [18]

B. Design and Implementation of Optical Character Recognition System to Recognize Gujarati Script using Template Matching:

Template matching, have been used wherein, each character in the input image as seen by OCR is compared against a set of templates and the code of the template that best matches is output.

Recognition Technique - A character is split into connected components and each component is then cut so as to remove the lower and upper modifiers from the glyph. They are matched against a database. These connected and cut components are called as OCR glyphs. [19]

To compute distance or dissimilarity between two templates, they should be of same size. So, all the glyph images are normalized to 32 × 32 size. The image of the input glyph is also scaled to 32 × 32 size before comparison. The method used to measure the similarity or distance between is crucial. The challenge in template matching is in making the matching process fast and robust against distortions.

C. Zone Identification In The Printed Gujarati Text:

This proposes a sophisticated method for accurate zone detection in images of printed Gujarati. We may think of two strategies for the recognition of the Gujarati Text: 1. recognizing the complete consonant-vowel cluster as a distinct symbol, or 2. first segmenting the consonants from a dependent vowel modifier and then recognizing them separately. [1]

If we take the first approach of recognizing the consonant-vowel cluster as a whole, then the number of glyphs to be identified increases enormously. This System uses an algorithm based on connected components and finding slope according that. [1]

D. Similar Looking Gujarati Printed Character Recognition Using Locality Preserving Projection And Artificial Neural Networks:

In this, Locality Preserving Projection (LPP) is used which is dimensionality reduction process that utilizes the redundancies present in the image. The projected images whose dimensionalities are drastically reduced by LPP are used directly as features. Specifically, an extended version of Supervised Locality Preserving Projection (ESLPP) has been utilized for dimensionality reduction of the character images. ESLPP preserves the local structure of the characters within a few most significant directions in the projection space and at the same time expected to increase the discriminating characteristics of the characters. [21]

Experiments are conducted on similar looking Gujarati characters to show the discriminating power of ESLPP coefficients used as features to a widely used BPNN classifier. BPNN is used as classifier in the present experimentation which has been widely used in various non-linear classification problems. Backpropagation algorithm is easy to implement and has become popular in pattern recognition problems. The algorithm has been implemented on similar looking Gujarati characters taking two or more characters at a time. A minimum efficiency of 96% recognition in all data sets is achieved.

V. Conclusion

In a typical OCR systems input characters are digitized by an optical scanner. In this paper we presented brief overview of each stage of OCR which is very initial attempt of character recognition. In this paper we have also presented survey of progress in Gujarati Character recognition.

Many applications are awaiting the enhancement in character recognition to be adopted it fully. The progress can extend by focusing on recognition of conjuncts of Gujarati Language.

References

[1] S.Rama Mohan, Jignesh Dholakia, Atul Negi. "Zone identification in the printed Gujarati Text", Processing of the 2005 Eight International Conference on Document Analysis & Recognition (ICDAR'05)

[2] R. M. K. Sinha, "A Journey from Indian Scripts Processing to Indian Language Processing", IEEE Annals of the History of Computing, pp8-31, Jan–Mar 2009.
[3] U. Pal, B. B. Chaudhuri, "Indian Script Character recognition: A survey", Pattern Recognition, vol. 37, pp. 1887-1899, 2004
[4] Vikas J Dongre Vijay H Mankar "A Review of Research on Devnagari Character Recognition" International Journal of Computer Applications (0975 8887) Volume 12– No.2, November 2010
[5] Rinku Patel, Avani Dave, "A Survey Paper on Character Recognition" IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 1 Issue 2, April 2014.
[6] D. Trier, A. K. Jain, T. Taxt, "Feature Extraction Method for Character Recognition - A Survey", Pattern recognition, vol.29, no.4, pp.641-662, 1996.
[7] R. Jayadevan, Satish R. Kolhe, Pradeep M. Patil, and Umapada Pal "Offline Recognition of Devanagari Script: A Survey", IEEE Transactions On Systems, Man, And Cybernetics—Part C: Applications And Reviews, Vol. 41, No. 6, November 2011
[8] B. B Chaudhuri and U. Pal, "AnOCR system to read two Indian language scripts: Bangla and Devanagari," in Proc. 4th Conf. Document Anal. Recognit., 1997, pp. 1011–1015.
[9] V.K.Govindan and A. P. Shivaprasad, "Character recognition:Asurvey," Pattern Recognit., vol. 23, pp. 671–683, 1990.
[10] N. Arica and F. T. Y. Vural, "An overview of character recognition focused on off-line handwriting," IEEE Trans. Syst., Man, Cybern. C: Appl. Rev., vol. 31, no. 2, pp. 216–233, May 2001.
[11] U. Pal and B. B. Chaudhuri, "Printed Devnagari script OCR system," Vivek, vol. 10, pp. 12–24, 1997.
[12] S. Kompalli, S. Nayak, S. Setlur, and V. Govindaraju, "Challenges in OCR of Devanagari documents," in Proc. 8th Conf. Document Anal. Recognit., 2005, pp. 1–5.
[13] V. Bansal and R.M. K. Sinha, "A complete OCR for printed Hindi text in Devanagari script," in Proc. 6th Conf. Document Anal. Recognit., 2001, pp. 800–804.
[14] Mitul Modi, Fedrik Macwan, Ravindra Prajapati, "Gujarati Character Identification: A Survey", International Journal Of Innovative Research In Electrical, Electronics, Instrumentation And Control Engineering Vol. 2, Issue 2, February 2014
[15] Vikas J Dongre Vijay H Mankar "A Review of Research on Devnagari Character Recognition" International Journal of Computer Applications (0975 8887) Volume 12– No.2, November 2010
[16] Patel C. & Desai A. , "Segmentation of Text Lines into Words for Gujarati Handwritten Text", International Conference on Signal & Image processing, 130–134.
[17] X. He and P. Niyogi, "Locality Preserving Projections", Proc. Conf. Advances in Neural Information Processing Systems, 2003.
[18] Antani S. and Agnihotri L. "Gujarati Character Recognition", In Proc. Of 5th International Conference on Document Analysis and Recognition, IEEE Computer Society Press, pp. 418-421, 1999.
[19] S K Shah and A Sharma "Design and Implementation of Optical Character Recognition System to Recognize Gujarati Script using Template Matching", IE(I) Journal-ET, Vol. 86, pp. 44-49, 2006.
[20] Mandar Chaudhary, Gitam Shikkenawis, Suman K. Mitra, Mukesh Goswami, "Similar looking Gujarati printed character recognition using Locality Preserving Projection and Artificial Neural Networks",Third International Conference on Emerging Applications of Information Technology (EAIT), 2012
[21] Line Eikvil "OCR-Optical Character Recognition", December 1993
[22] Hetal R. Thaker, Dr. C. K. Kumbharana, "Analysis of structural features and classification of Gujarati consonants for offline character recognition", International Journal of Scientific and Research Publications, Volume 4, Issue 8, August 2014 ISSN 2250-3153

