

A Review of Handwritten Character Recognition

Nikita Mehta

PhD Scholar, GLS University,
Ahmedabad
Gujarat

Jyotika Doshi, PhD

Associate Prof., GLS(SRP)ICT
Ahmedabad
Gujarat

ABSTRACT

The aim behind Optical Character Recognition is to create human like perception and character identification by artificial systems. A lot of work has been done for printed and handwritten character recognition for many languages across the world. Even for many Indian languages, a good amount of work is done, but it could not get that accuracy as English, German etc. languages because of its complexities. In this paper various techniques for Handwritten Character Recognition (HCR) are reviewed and analyzed.

General Terms

Natural Language Processing, Pattern Recognition, Optical Character Recognition et. al.

Keywords

Optical Character Recognition (OCR), Handwritten Character Recognition (HCR), Binarization, Segmentation, Feature extraction.

1. INTRODUCTION

Optical Character Recognition is a popular area of research in current era of technology. Basically, it is a machine learning process in which a machine is made to read like human.

An OCR is helpful in many ways. It can be used to preserve the literature (ex. ancient manuscripts) which is on paper but not digitized, to recognize vehicle number from CCTV images and so on. For OCR activities, a lot of work has been done and research is still going on for better results.

On the basis of data acquisition techniques, OCR activity can be divided in two major categories: Online Optical Character Recognition and Offline Optical Character Recognition [8]. Offline method can further be divided in two sub categories – Printed and Handwritten.

Online Character Recognition:

It is dynamic and real time system. It processes character while it is being created. External factors like position of pen, pressure, speed of writing, stroke making etc. will play an important role in character identification [7]. Details related to pen movement for making curves and strokes will also help in this process. Identification of text written using stylus on handheld devices is an example of online character recognition.

Offline Character Recognition:

In Offline, existing document is digitized, stored and then processed [7]. Thus, it does not have any external information, thus making the recognition process difficult. Offline recognition may be used for printed or handwritten documents [8].

- Printed Document: The positive factor for printed material is having characters of uniform style and size. This makes it easier to recognize.
- Handwritten Documents: Handwritten character recognition (HCR) system is very hard and complex. Written characters differ from writer to writer. Even a single person's written characters may differ at different point of time within a word, sentence, paragraph or document [17].

2. CHARACTERISTICS OF DEVNAGARI FAMILY SCRIPT

Devnagari script is derived from Nagari (inherent of ancient Bramhi script which is parent script of most of Indian scripts and some of Asian scripts) script. Devnagari was originally developed for Sanskrit language; but then it was accepted to write many Indian languages like Hindi, Marathi, Nepali, Bhojpuri, Marwari, etc. [8]. Most of other North Indian scripts like Gujarati and Punjabi are inherited from Devnagari script.

Like other Indian scripts, Devnagari script is quite complex in structure. It has 33 consonants and 13 vowels. Devnagari characters have 'Shirorekha' above all characters. Other than these basic characters, script uses some modifiers in combination with consonants. These modifiers are placed in above, below, right and left of the character. The script also uses joint characters or conjunction characters at many places. All these issues make character recognition process harder [1].

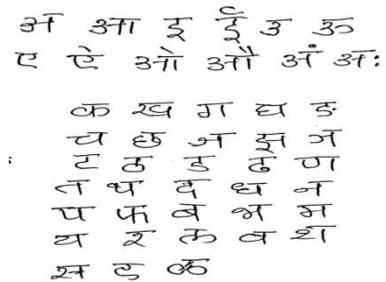


Fig 1 : Devnagari Vowels and Consonants

Gujarati script is derived from Devnagari script. It is a variant of Devnagari as it does not have 'Shirorekha' above the characters but, it do have modifiers and joint characters like Devnagari.

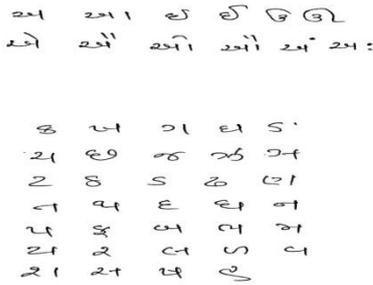


Fig 2 : Gujarati Vowels and Consonants

3. STEPS FOR OCR PROCESS

Following are major steps of optical character recognition process:

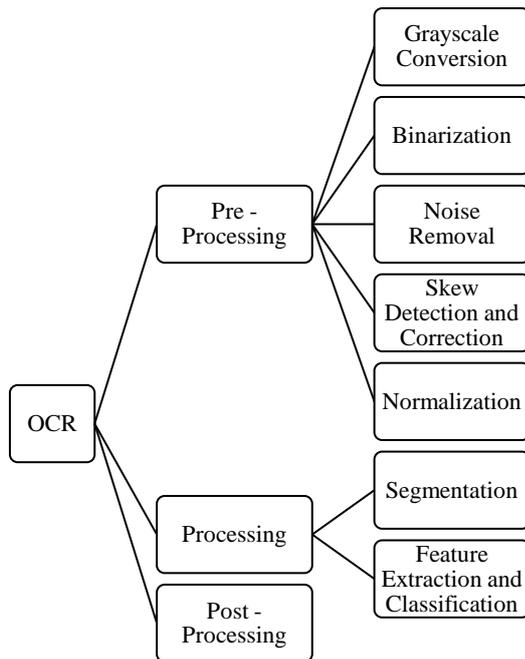


Fig 3 : OCR Process

3.1 Pre – Processing

Input of OCR is a scanned document as an image and the image is processed for character recognition. Before the final stage of actual character recognition, image needs to be pre-processed to make it better for more accurate results. The image pre-processing involves:

3.1.1 Grayscale Conversion

Scanned images will be in RGB format. For OCR process to get accurate result, image need to be converted in Grayscale format. There are several mathematical formulas which can generate a gray value for any pixel with RGB values. Some sample formulas are [22]:

- gray = red
- gray = green
- gray = blue
- gray = (red + green + blue)/3

3.1.2 Binarization

For grayscale images with information such as text or graphics, binarization is usually performed first. The objective of binarization is to automatically choose a threshold that

separates the foreground and background information [11] [15].

There are two approaches for conversion of gray level image to binary form: 1. local or adaptive threshold, 2.global threshold [8] [3] [2].

3.1.2.1 Local Threshold

Local or adaptive threshold uses different values for each pixel according to the local area information [8]. In degraded document images, where considerable background noise or variation in contrast and illumination exists, there exist many pixels that cannot be easily classified as foreground or background. In such cases, binarization with local thresholding is more appropriate [18].

The techniques of Bernsen, Chow and Kaneko, Eikvil, Mardia and Hainsworth, Niblack, Taxt, Yanowitz and Bruckstein, and Sauvola and Pietikainen are some examples of local threshold [18].

3.1.2.2 Global Threshold

Global threshold selects single value based on estimation of the background level from the intensity histogram of the image. In document like images with uniform contrast distribution of background and foreground, global thresholding is more appropriate [18].

Some global threshold techniques are: SIS Threshold, Otsu Threshold and Iterative Threshold.

3.1.3 Grayscale Conversion

Digital image contains variety of noises because of degradation of documents due aging, moisture, fungus, photocopying etc. [4] [15]. This noise can lead to less precise character recognition. To avoid this, noises should be removed from images.

Filters like: Mean filter, Median filter, Gaussian blur filter, etc. are used for removing noises from images [4].

3.1.4 Skew Detection and Correction

Skewness may be there due to many reasons like writing style or paper feeding style. Handwritten documents may be originally skewed because of a person's writing style. Skew might be introduced while feeding the document to the scanner; ex. document is not properly aligned while scanning.

Skew detection technique determines the angle by which the document is skewed while it was digitized. For document skew detection - Projection profile, Hough transform etc. techniques can be used [19].

After skew detection in document, the next task is to correct that skew angle. Some techniques used for skew correction are: projection profile technique, linear regression analysis, fourier transform based method, nearest neighbor chain, edge based connected component approach, interline cross-correlation, entropy based methods [3].

3.1.5 Normalization

Handwritten characters will be of different sizes, slant and rotation according to writing styles of different persons. To make characters of uniform size, slant and rotation; normalization is applied [3]. Normalized character images will make recognition rate higher than normal images.

Some techniques which could be used for Normalization are: Linear normalization, Moment normalization, Slant normalization and Non-linear normalization.

3.2 Processing

After noise removal and size normalization, an image is ready for character recognition. In this processing phase, various techniques are used to identify actual character.

3.2.1 Segmentation

Segmentation is an important part of Optical Character Recognition. In this process, if document contains both text as well as graphics, then they are separated in the first pass. After separating, only text part is to be processed. Text part is sub divided horizontally using row histogram to separate all horizontal lines. Then using vertical histogram, words are separated from line and then individual characters get separated [11] [7] [15].

Desai, A. A [10] presented radon transformation for document segmentation. Radon transform computes line integral from multiple line sources along a parallel path in a direction. It computes projections of an image from a specific direction, taken from different angles by rotating the source of projection around centre of image. This technique will extract lines even from noisy images.

Accuracy of final character recognition depends on quality of segmentation.

3.2.2 Feature Extraction and Classification

To identify the character separated in Segmentation phase, it requires matching it with all the characters in the database. This is quite time consuming and processor consuming process.

To make the processing more efficient, characters of similar characteristics are grouped together as a set. In [14] authors suggested that Devnagari characters can be classified into three categories based on the presence of the vertical bars. This could be the characters with end-bar, the characters with middle bar and the characters without any bar line.

All characters have some specific characteristics or features using which one can identify the character. Features includes: Number of horizontal lines, Number of vertical lines, Number of right diagonal lines, Number of left diagonal lines, etc. Feature Extraction is a process to extract these specific characteristics [16].

In [6], authors created seven sets of Gujarati characters using primary features with tree classifiers: number of objects in the character, number of objects in upper half of a character, number of objects in lower half of a character, number of holes into character.

Desai, A. A [1] mentioned that there are three types of feature extraction methods are there: structural, statistical and hybrid. He compared results of all three methods and found hybrid feature set more effective than other two for Gujarati alphabets.

Authors of [14] mentioned box approach for feature extraction. Here, character image is divided in 24 equal sized boxes. Thereafter, all the boxes having image portions as well as empty boxes are analyzed.

After feature extraction, next step is to identify the character. Number of character identification techniques are proposed and used by different researchers for the same. Some of them are Template matching, Nearest Neighbor Classifier (NNC), KNN (K Nearest Neighbor), Kohonen SOM etc. [9] [13].

3.3 Post – Processing

Cleaning an image in pre-processing stage may cause some important information to be lost. This may result into improper segmentation leading to erroneous recognition of character. For a more precise OCR, these errors need to be identified and corrected. A simplest way to achieve this is to apply a spell check on resultant document. But this is quite difficult for languages other than English, especially for Indian languages. This requires extensive research [21].

Handwritten Character Recognition process is a feed forward or back propagation method. It requires teaching machine every time an unidentified or a new character encounters. [13]

4. CONCLUSION

Character Recognition is one of the popular areas of research nowadays. This paper gives overview of various existing approaches for Handwritten Character Recognition. Many researchers have achieved good recognition rate for Indian scripts, but this rate is for an individual characters. Character with different modifiers and specially connected or joint characters are quite complex to identify. So, a lot more research is needed for Indian scripts to be identified accurately with all its complexities.

5. REFERENCES

- [1] Desai, A. A. (2015). Support vector machine for identification of handwritten Gujarati alphabets using hybrid feature space. *CSI Transactions on ICT*, 2(4), 235-241.
- [2] Thaker, H. R., & Kumbharana, C. K. (2014). Structural Feature Extraction to recognize some of the Offline Isolated Handwritten Gujarati Characters using Decision Tree Classifier. *International Journal of Computer Applications*, 99(15), 46-50.
- [3] Modi, M., Macwan, F., & Prajapati, R. Gujarati Character Identification: A Survey
- [4] Magare, S. S., Gedam, Y. K., Randhave, D. S., & Deshmukh, R. R. (2014). Character Recognition of Gujarati and Devnagari Script: A Review. *International Journal of Engineering*, 3(1).
- [5] Patel, C., & Desai, A. (2013). Extraction of Characters and Modifiers from Handwritten Gujarati Words. *International Journal of Computer Applications*, 73(3).
- [6] Patel, C., & Desai, A. (2013, June). Gujarati handwritten character recognition using hybrid method based on binary tree-classifier and k-nearest neighbour. In *International Journal of Engineering Research and Technology* (Vol. 2, No. 6 (June-2013)). ESRSA Publications.
- [7] Shah, M., & Jethava, G. B. (2013). A literature review on hand written character recognition. *Indian streams research journal*, 3(2), 1-19.
- [8] Indira, B., Qureshi, M. S., Shaik, M. S., Saqib, R. M., & Murthy, M. R. (2012). Devanagari Character Recognition: A Short Review. *International Journal of Computer Applications*, 59(6).
- [9] Sojitra, B., & Dhakad, V. (2012). Neural Network in Character Recognition of Gujarati Script. *Journal of Information, Knowledge and Research in Computer Engineering*, 2(2), 269-272.

- [10] Desai, A. A. (2012). Segmentation of characters from old typewritten documents using radon transform. *Int. J. Comput. Appl.*, 37(9), 0975-8887.
- [11] Maloo, M., & Kale, K. V. (2011). Gujarati script recognition: a review. *International Journal of Computer Science Issues*, 8(4).
- [12] Singh, B., Mittal, A., Ansari, M. A., & Ghosh, D. (2011). Handwritten Devanagari Word Recognition: A Curvelet Transform Based Approach. *International Journal on Computer Science and Engineering*, 3(4), 1658-1665.
- [13] Dongre, V. J., & Mankar, V. H. (2011). A review of research on Devnagari character recognition. *arXiv preprint arXiv:1101.2491*.
- [14] Hanmandlu, M., Murthy, O. R., & Madasu, V. K. (2007, December). Fuzzy Model based recognition of handwritten Hindi characters. In *Digital Image Computing Techniques and Applications, 9th Biennial Conference of the Australian Pattern Recognition Society on* (pp. 454-461). IEEE.
- [15] Kasturi, R., O'gorman, L., & Govindaraju, V. (2002). Document image analysis: A primer. *Sadhana*, 27(1), 3-22.
- [16] Dileep, D. (2012). A feature extraction technique based on character geometry for character recognition. *Department of Electronics and Communication Engineering, Amrita School of Engineering, Kollam, India*.
- [17] <http://stackoverflow.com/questions/18620977/best-setting-for-scanners-for-scanning-documentstiff-and-pdf>
- [18] Singh, T. R., Roy, S., Singh, O. I., Sinam, T., & Singh, K. (2012). A new local adaptive thresholding technique in binarization. *arXiv preprint arXiv:1201.5227*.
- [19] Gatos, B., Papamarkos, N., & Chamzas, C. (1997). Skew detection and text line position determination in digitized documents. *Pattern Recognition*, 30(9), 1505-1519.
- [20] Liu, C. L., Nakashima, K., Sako, H., & Fujisawa, H. (2004). Handwritten digit recognition: investigation of normalization and feature extraction techniques. *Pattern Recognition*, 37(2), 265-279.
- [21] Arica, N., & Yarman-Vural, F. T. (2001). An overview of character recognition focused on off-line handwriting. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 31(2), 216-233.
- [22] Seven grayscale conversion algorithms (with pseudocode and VB6 source code) <http://www.tannerhelland.com/3643/grayscale-image-algorithm-vb6/>