International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320-088X

IMPACT FACTOR: 7.056

IJCSMC, Vol. 10, Issue. 9, September 2021, pg.12 – 22

Script Identification for Devanagari and Gurumukhi using OCR

Shubhankar Sharma¹; Vatsala Arora²

Department of Computer Science and Engineering, Amity University, India

¹shubhsharma2905@gmail.com; ²arora.vatsala09@gmail.com

DOI: 10.47760/ijcsmc.2021.v10i09.002

Abstract — The study of character research is an active area for research as it pertains a lot of challenges. Various pattern recognition techniques are being used every day. As there are so many writing styles available, development of OCR (Optical Character Recognition) for handwritten text is difficult. Therefore, several measures have to be taken to improve the recognition process so that the burden of computation can be decreased and the accuracy for pattern recognition can be increased. The main objective of this review was to recognize and analyze handwritten document images. In this paper, we present a scheme to identify different Indian scripts like Devanagari and Gurumukhi.

Keywords— optical character recognition, gurumukhi, devanagari, segmentation, script, image, graphics, text detection, natural language processing, script Identification

I. INTRODUCTION

The main purpose of this optical character recognition is the identification of scripts Gurumukhi and Devanagari. If the monolingual OCR works, OCR always treat these scripts as waste and therefore these scripts remain unidentified. This OCR helps identify these scripts so that they can be identified, and the person can understand the script used in the given text. This OCR used in a variety of ways, e.g. by identifying the script used in the given digital image and by processing, the text can be identified. Then the user can translate the text in the most appropriate language. Because this OCR identifies the Gurumukhi and Devanagari scripts, it must be programmed to handle both scripts. Both scripts, such as Gurumukhi and Devanagari, divided into three horizontal parts that hinder the development of OCR. OCR technology is extremely useful because it saves time by avoiding rewriting consumption. The digital image simply scanned, processed and displayed in editable text.

India is a multilingual country with different screenplays. In the last fifty years, optical character recognition systems (OCR) have come a long way from unique specialized readers to current multi-use production and interactive online systems. Therefore, the development of a successful multilingual OCR system for extracting functionality from different scripts is a very important step. This advancement has reduced the cost of data collection and led to the development of more reliable and accurate OCR systems. Family procedures to measure relevant shape information in a model to facilitate model classification called feature extraction. Commercial OCR systems for Latin characters are widely used in personal computers. In addition, systems that are available commercially can read different writing styles (for example, Omni printed handsets) and fonts, including Chinese, Japanese, Korean, Criminal and Arabic. Modern OCR software is extremely precise, easy to use and convenient. For the first time, it appears that OCR used in all types of mass-market work environments. Analyze a segment of text and select features that are unique to text and can be used for identification. The soul of the model recognition system is the selection of characteristics that are stable and representative.

II. GOALS AND OBJECTIVES

Our proposed system is a visual feature Characterization (OCR) in a network i.e. a credential that supports different personal identities with different scripts. This feature called infrastructure compartments, which eliminates the problem of defining different characters and supports a number of features that run in the document. Some features include editing and searching, as the current system only supports document modifications. In this regard, the framework refers to the development of a structure that supports a particular set of languages. Our focus and goal are to collect the data from the Gurmukhi script and Devanagari script and to binarize the image, which is the editing of the 0 and 1 format images that the system reads. There is no OCR system sufficient for recognizing handwritten texts. This is because characters not automatically separated in handwritten text images. In order to translate this kind of textual image, holistic approaches are called "disconnected HTR segmentation" to solve the problem. In these methods, all elements of the text (expressions, words and trade unions) must be fully recognized, without the image of these elements being folding ten times. HTR's latest technology has provided ideas and methods for automatic rhetoric (ASR), such as the hidden Markov model and N-grams. For this technology, images of handwritten documents need reprocessing for document images, such as building analysis and text execution. Because of the large amount of data, these steps should be automating as much as possible. This is a challenge because the standard rules for handwritten documents are missing. When the text line images are available, the HTR model can be obtained automatically with a strong and well-known training.

III. METHODOLOGY

The main methodology that is to be followed was to identify the script that was inputted in the system. The image that is inputted has to undergo the processes of OCR. The OCR performs its different process on the image so that the script in the image is identified.

Fig. I – Block Diagram of OCR

Pre-Processing:

The pre-processing phase includes various operations that are sequentially applied to the input image for effective character recognition. Digital image processing uses computer algorithm to create, work, share and publish digital images.

Sample analysis is an interdisciplinary area built in other fields, such as advanced processing, artificial vision and artificial intelligence. Thus, OCR cannot be used without image processing and / or information gathering. Each OCR system runs at different levels, including data collection, preprocessing, Segmentation, Extraction, and post-processing, but the most important is the information management needed to convert data into ongoing errors collections. Add a cradle restriction or create data for subsequent activities in the description or grading phase. Data processing describes all processing that has been handled for other raw materials. Therefore, preprocessing is a preparatory step that converts data to formats that can be done more effortlessly and effectively. Therefore, the main objective of working on data is mainly to reduce change, leading to an abatement in the frequency of accrual & complexity. For example, pre-treatment of raw materials is essential to the success of efficient personal identification systems. Before you can check if this is OK, pre-treatment is the necessary step in extracting. The step pattern system is in queue, which means that each step depends on the performance of past chats for achieving maximum / valid outcomes. Nonetheless, it is obvious that attribute weights that best match to the grading level are only created by facilitating the pre-processing phase. The pre-treatment is primarily aimed at standardizing and preventing deficiencies that would otherwise make classification more difficult and reduce recognition speed.

The uses of Digital image processing algorithms:

- It can easily convert images from image detector to digital images.
- Improve picture clarity, quality and eliminate noise and other sectors
- It achieves the size, scale or number of objects in sight
- Prepare the images to be displayed or printed and compress the image

Segmentation:

The segmentation phase is an important phase of OCR, since poor segmentation can lead to false detections. It has a great impact on the overall accuracy of the recognition process, since poor segmentation can lead to erroneous recognition. When analyzing the segmentation of scripts, the segmentation essentially comprises the line, the word and the character. The process of Segmentation is of breaking down of an image / document to separate and homogeneous areas. This task is solved by finding the limits. There are several approaches to determining character limits. The Devanagari document is broken down into sequences of lines and words with horizontal and vertical projection. The words of Devanagari can be further subdivided by eliminating the Shiro-Rekha. A word of Devanagari and Gurumukhi can be divided into three parts. The main characters are in the central part. The upper part shows the area on Shiro-Rekha and the optional modifiers can be found at the bottom. Therefore, the segmentation of the Devanagari and Gurumukhi figures is due to several modifiers.

Observation segmentation is considered one of the most important choices, including curious, Arabic, Urdu, and other scripts. Then, many techniques are developed for this segment, and most of them are language-specific and functionally functional with other languages. Viewer segmentation requires only a handwritten manuscript document based on a handwritten view. The printed latex seams are smooth with horizontal and vertical histograms. Smaller and more solo scripts that contain serifs can cause problems requiring additional treatment to solve the problem of bullying.

Once the words are divided into lines of text, use the method described here. The segmentation process is applied so that each character leaves segmented words. Headers or ribbon identifies and deletes before drawing-level word segmentation. This is done by placing the lines with the max number of black dots in the word, using a heuristic approach to find the shield, because not every Shirorekha lines contain the same number of black pixels. We removed the black dot in this document. The difference is calculated as 10 between the Shi-rorekha rows, although it is not the same for all other scenario. This representation of differences for Hindi was selected after a thorough analysis of many Shi-rorekha lines. Once Shirorekha is located, it becomes white pixels.

When Shiro-rekha is removed, the word is divided into three horizontal zones, i.e. upper, middle & lower zones. Single characters are divided by vertical scanning of each area. It is performed only for area width and word length. Only modifiers are present in the upper and lower regions. Therefore, prior to vertical exploration in these areas, whether modifiers are present or not. This is necessary because in some cases words can only have words from the central region or central and upper regions or central and lower regions.

The horizontal profile method is used to calculate the sum of all white points on each line. It gives a corresponding histogram of this image. The steps in the line segment are the following:

Line Segmentation:

1. Create a horizontal histogram of the image.
2. Look for the threshold value corresponding to the difference between the corresponding histograms.
3. Define each histogram by reference value and save it.
4. We get broken lines of the picture.

Fig. II– Example of Line Segmentation

Word Segmentation:

In syntax, we use a vertical projection format method to calculate the sum of all white points. Draw the name of the calculated white dots. The steps in the division of the words are as follows:

1. Create the vertical histogram of the image.
2. Find the threshold value
3. Separate the corresponding histogram and save it
4. Recorded images are segmented words in line
5. Repeat the process for each line

Fig. III– Example of Word Segmentation

Character Segmentation:

In the nature of the division, we use horizontal and vertical projection profiles methods to understand characters from words. The steps for spatial division are as follows:

1. Develop a horizontal histogram of the image.
2. Draw white chart's histogram account.
3. Look for the header line or the shirorekha and remove it.
4. Create a vertical histogram
5. Find threshold values and separate histogram
6. Save part image as characters

Fig. IV– Example of Character Segmentation

Fig. V– Thinning

Thinning:

It is another process analogous to skeletonization that reduces the width of the binary object to a single pixel. Clears the selected foreground pixels of the binary image. By diluting the image, a skeleton is extracted from the image without losing the structural properties of the image. The dilution code consists of edge pixel analysis & connectivity analysis.

It is a morphological process that first removes pixels, such as erosion or aperture from a binary image. It is used in some application; it is especially required for skeletonization. It is typically used to adjust the output of the edge detector by decreasing all of the thickness of a single pixel. Dilution is only applied to binary pictures and produces a binary pictures that is different from the output.

Digital image processing includes three words: digital, picture, and processing. It works with pictures that capture a digital image, because the input side reduces noise and signal distortion and provides the result. It works for the system and the signals, but mainly for the images. It uses an efficient algorithm to give an image as a production. We will compare this problem to that of your choice. If it does not comply with this edition, we will treat it until we have received our product as desired or approximately. The processing of digital images is much better than analog signals because it does not work with multidimensional pixels. The digital image is a group or matrix defined by a no. of bits.

Representation of the structural form of digital images is one of the basic conditions. This can be achieved by reducing in a graph. This reduction can be achieved by taking the skeleton of the region using skeletonization, also known as skipping. The process of extracting the outlines of an object in a given digital picture is called skeletonization. The black dots of the voice are extended to the skeleton of the width in pixels is a morphological operation. Dilution is the first step we can call "pre-treatment". We can redefine skeletal thinning by extracting or reducing a digital image to its minimum size, or by reducing the image so that the image retains the need for image processing points. We can evaluate the performance using the following parameters: - Processing time, connection, image must not contain any noise and no branching. Although many practices were taken to implement the rejected algorithm, they could not meet the essential requirements. Sound sensitivity: - This means that the image must not contain any pixel resulting in an incorrect output. We can say that if noise occurs after using the algorithm, the desired output cannot be obtained. Recoverable: - If we apply an algorithm to the diluted image, we should use the production as a diluted image or original image. Skeletonization is generally applied to binary images composed of black (first) and white pixels. A binary image is taken which produces another binary image.

Fig. VI– Tail generation at acute limb angle

Dataset collection:

Data-set (or dataset) is a set of various different types of data which is collected from different sources (like through sensors, Surveys online applications communicated through Internet, etc). Generally, the data set is a database table or amount of statistical data, with each column of the table representing a particular variable, and each row corresponding to a particular table in the table data. The data shows the value of each variable, such as the length of data (i.e. string length or length of numerical value) and weightage (i.e. how much critical the data is, what is the density or flow of data coming in, etc.) of the object, for each member of the data package, each value is called data. The data may contain data for one or more members, corresponding to the number of rows.

Skew Detection:

While the document is scanned, you can flip the entire document or part of it, if it can be loaded through the page scanner. Some pages may not be inserted directly into the scanner, but this tends to skew the bitmap images of those pages. In this way, document skewing often occurs when scanning or copying documents. These effects appear visually and tilt the text line in conjunction with the x axis and mainly affect the alignment of all the text lines. By applying the principles of trigonometry, the images can be corrected accordingly.

Slant Detection:

The inclination of the character normally present in the script is called tilt. The following figure shows some examples of inclined hand numerical strings. The correction of the inclination is an important step in the preprocessing phase of handwritten words and the recognition of numerical strings. The general purpose of a slope correction is to reduce text changes, especially to improve the quality of candidates for the division of words or numbers into a string, which can provide better accuracy recognition.

Factors affecting Character Recognition Quality:

Text recognition accuracy is affected by many factors of OCR. Such factors majorly consist of quality of scanner, scanner dpi resolution, document category printing (laser-printed/ photocopy), sheet quality, text types used in the text, language complexity & dictionary used. The terms "withdrawal" and "visual text" in old sheet documents, watermarks & incoherent descriptions are samples of problems that affect the accuracy in the OCR on white

background. For example, shows the image of the enlightened gray color document. Below figure shows a composite background image composite file. Other features include printing operations, such as consistency, text account and appearance and graphics.

Fig. VII– Composite background image

Post Processing:

Choosing a stable and representative group is at the heart of the model recognition system. We have adopted a multilevel post-processing technique.

- Error detection.
- Error correction.
- Language model
- High channel model

Error Detection Model:

This model is used to determine the number of errors and to specify the position in which it occurs. Here we analyze the image, first we scan the most common characters of Devanagari and Gurumukhi. And then we analyze the number of correct candidates with the number of errors. Using this relationship, we can find a threshold to detect the point where the error occurred for the first time.

Error Correction Model:

- Linguistic model

From the error detection model, we know where the error occurred. Therefore, we use a segmentation model to correct these errors. This approach is very useful and effective in correcting the error in long words made up of two or more characters. From the result of the experiment, the correction state eliminates two-thirds of the error generated by the OCR.

- Channel models of noisy models.

It may be that some correct characters are missing from our list of candidates. Therefore, we use a noisy-channel model to add few characters to list of candidates to solve this conflict. Therefore, this model is also very useful for eliminating errors. Therefore, the post-processing technique helps to find errors and correct errors. This phase allows us to exit without errors.

IV.RESULTS

Fig. VIII– Read Image

Fig. IX– Image converted to grayscale

Fig. X– Blobs analysis of image

Fig. XI– Plotting the object location

V. CONCLUSION

We have worked on the processes of OCR (Optical character Recognition) for Indian script language Gurumukhi and Devanagari which compromise the recognition accuracy of different language scripts. As we know India has multiple languages, so we have distinguished between two Indian script languages i.e. Devanagari & Gurmukhi. Most of the project work has been done on MATLAB. First we collected data set of Hindi text. Then our system performs various operations on uploaded image.

First of all, preprocessing is performed on uploaded image, then it processes & filters the image, remove noise and make image clear for further processing. After pre-processing our system perform segmentation, this separate and read each word from uploaded image. Then we have tried different things with three unique methodologies for content characterization. The customary methodology, first distinguishing the content of the word picture and after that sending it for acknowledgment, then it is compared with our data set and it tell the accuracy that is how much percentage of the given text is in Gurmukhi or Devanagari or in some other text.

ACKNOWLEDGEMENT

This research paper came to us as an excellent opportunity and gave us such an exposure. We would like to thank the authorities of **AMITY UNIVERSITY** and our faculties to providing us the appropriate knowledge and skills which was the encouragement for this paper. We would also like to thank **IJCSMC JOURNAL** for their support in publishing this paper.

REFERENCES

- [1]. Joshi G.D., Garg S., Sivaswamy J. (2006) Script Identification from Indian Documents. In: Bunke H., Spitz A.L. (eds) Document Analysis Systems VII. DAS 2006. Lecture Notes in Computer Science, vol 3872. Springer, Berlin, Heidelberg. https://doi.org/10.1007/11669487_23
- [2]. Reya Sharma, Baij Nath Kaushik and Naveen Kumar Gondhi, 2018. Devanagari and Gurmukhi Script Recognition in the Context of Machine Learning Classifiers. *Journal of Artificial Intelligence*, 11: 65-70
- [3]. Joshi, Gopal & Garg, Saurabh & Sivaswamy, Jayanthi. (2007). A generalised framework for script identification. *IJDAR*. 10. 55-68. [10.1007/s10032-007-0043-3](https://doi.org/10.1007/s10032-007-0043-3).
- [4]. Kurban Ubul, Gulzira Tursun, Alimjan Aysa, Donato Impedovo, Giuseppe Pirlo, Tuergeren Yibulayin, "Script Identification of Multi-Script Documents: A Survey", *Access IEEE*, vol. 5, pp. 6546-6559, 2017

- [5]. N. Shobha Rani, M. Keerthi, "Wavelet Analysis based Handwritten Script Classification using Auto Encoders", *Intelligent Computing and Control Systems (ICICCS) 2020 4th International Conference on*, pp. 607-614, 2020.
- [6]. Sk Md Obaidullah, K. C. Santosh, Chayan Halder, Nibaran Das, Kaushik Roy, "Automatic Indic script identification from handwritten documents: page block line and word-level approach", *International Journal of Machine Learning and Cybernetics*, 2017
- [7]. Soumya Ukil, Swarnendu Ghosh, Sk Md Obaidullah, K. C. Santosh, Kaushik Roy, Nibaran Das, "Improved word-level handwritten Indic script identification by integrating small convolutional neural networks", *Neural Computing and Applications*, 2019
- [8]. Soumyadeep Kundu, Sayantan Paul, Pawan Kumar Singh, Ram Sarkar, Mita Nasipuri, "Understanding NFC-Net: a deep learning approach to word-level handwritten Indic script recognition", *Neural Computing and Applications*, 2019
- [9]. C. Sunil, K. S. Raghunandan, H. K. Chethan, G. Hemantha Kumar, *Recent Trends in Image Processing and Pattern Recognition*, vol. 1037, pp. 72, 2019
- [10]. Iman Chatterjee, Manosij Ghosh, Pawan Kumar Singh, Ram Sarkar, Mita Nasipuri, "A clustering-based feature selection framework for handwritten Indic script classification", *Expert Systems*, pp. e12459, 2019
- [11]. Pawan Kumar Singh, Ram Sarkar, Vikrant Bhateja, Mita Nasipuri, "A comprehensive handwritten Indic script recognition system: a tree-based approach", *Journal of Ambient Intelligence and Humanized Computing*, 2018
- [12]. Nouf M. Alzahrani, Adil F. Alharthi, "Textural Measure for Medical Words Characterization Applied to Script Identification in Bilingual Context", *Augmented Human Research*, vol. 5, 2020
- [13]. Soumya Ukil, Swarnendu Ghosh, Sk Md Obaidullah, K. C. Santosh, Kaushik Roy, Nibaran Das, *Recent Trends in Image Processing and Pattern Recognition*, vol. 1380, pp. 499, 2021
- [14]. Pawan Kumar Singh, Iman Chatterjee, Ram Sarkar, Elisa Barney Smith, Mita Nasipuri, "A new feature extraction approach for script invariant handwritten numeral recognition", *Expert Systems*, vol. 38, 2021