International Journal of Engineering and Designing Innovation (IJEDI) ISSN(O): 2582-0788
Monthly, Peer-Reviewed, Refereed, Indexed Journal Impact Factor : 5.124
Volume - 6, Issue – 6 , June - 2024

DOIs:10.2019/IJEDI/202406001 --:-- Research Paper / Article / Review

An Enhanced method to recognize the optical characters

Ms. Supriya Lalasaheb Sukale
Research Scholar, Dept. of Computer Science and Engg.
Dr.BAMU Sub campus, Osmanabad
Email: supriyasukale14@gmail.com

Abstract: *Optical Character Recognition (OCR) is an interesting and challenging field of research in pattern recognition, artificial intelligence and machine vision and is used in many real life applications many times. The work done for the recognition of degraded Devanagari script is negligible in literature despite it is being used by millions people in India and abroad and it has numerous applications. Research on Optical Character Recognition OCR of degraded Devnagari script is very challenging due to the complex structural properties of the script that are not observed in most other scripts. Devnagari is the script for Marathi. The Marathi language contains 49 distinct characters, 12 vowels and 37 consonants. Recognition of Devnagari characters poses great challenge due to the large variety of symbols and their proximity in appearance and shape. Feature extraction and classification are the two very important steps in Optical character recognition for any language.*

Keywords *Digital image processing, OCR, Segmentation, classification.*

1. INTRODUCTION:

OCR is acronym for Optical Character Recognition in which text images are converted into digital text without human intervention and we converted it into editable format. OCR technology converts read only documents into digitized formats that can easily be retrieved, searched, and archived. Document analysis and recognition are two challenging research areas in pattern recognition. Although sufficient amount of research work is reported for printed offline OCR, little research work exists for offline handwritten OCR due to the diversified nature in handwritings. OCR usually involves three processes, namely text localization, character segmentation and recognition. The recognition is an important area of document image analysis, which is in mature stage for machine-printed text. However, for intermixed texts in multilingual environment, it still remains a challenging problem. Complications arise because of different writing styles and font sizes. For correctly recognizing a character, its segmentation plays an important role, as it is responsible for separating the characters from word images [1]. Sometimes multiple touched characters create problems during segmentation and consequently recognition error takes place. Moreover, Indic script like Devanagari contains vowels and consonants with modifiers, which make complicated compositions and present additional challenges for segmentation and recognition.

Recognition of Hindi characters is finished by utilizing a three stage system. Initial step is pre-processing, in which binarization of the picture and detachments of characters are performed. The following step is extraction in which either of the existing techniques of feature extraction is used. Third step is testing process.

The difficulty of performing accurate recognition is determined by the nature of the material to be read and by its quality. Generally, misrecognition rates for unconstrained material increase progressively from machine print to handwritten writing. Methods of increasing sophistication are being pursued. Current research employs models not only of characters, but also words and phrases, and even entire documents.

There is a huge amount of historical documents in libraries and in various National Archives that have not been converted digitally In handwritten OCR segmentation of text into characters is complicated task and which further reduces recognition accuracy. With respect to the segmentation of handwritten words into characters it is a critical task because of complexity of structural features and varieties in writing styles.

In the field of pattern recognition and artificial intelligence Optical Character Recognition (OCR) has become one of the most successful applications. Today, reasonably efficient and inexpensive OCR packages are commercially available to recognize printed texts in widely used languages such as English, Chinese, and Japanese. OCR is the most essential part of Document Analysis System that converts the scanned images of text , books, magazines, and

newspapers into machine-readable text. The process of Document Analysis Recognition can be divided into two parts, namely, printed and handwriting character recognition. The printed documents can further be divided into two parts: good quality printed documents and degraded printed documents. Degradation of the text leads to touching, broken, heavy printed, typewritten, faxed document.

It is used to decompose an image of a sequence of characters into sub images of individual symbols by segmenting lines and words. Segmentation is to separate line, word and character from the image. There are two types of contextual segmentation depending on signal discontinuity and signal similarity. Cluster, compression based methods, histograms, edge detection are widely used in contextual segmentation. Handwriting Recognition Technology has been improving much under the purview of pattern recognition and image processing since a few decades.

It is observed that most of the character segmentation and recognition methods reported in the literature are developed for specific script and text (either printed or handwritten). Furthermore, performance of Indic script OCRs is poorer than Latin script OCRs. Hence, the focus of the proposed research work is to segment and recognize the characters from scanned multilingual Indian documents with intermixed texts. The main contributions of the proposed research work are as follows:

For character segmentation, a hybrid scheme is pro- posed where at the beginning; pre-processing is per- formed on word images. Subsequently, one pixel width image, i.e. thinning operation is applied on words and projection profile is used to find primary segmentation paths. Later, distance related criteria are utilized to get fine segmentation paths. During post-processing step, over-segmented and overlapped characters are separated using graph distance theory. Finally, a trained Support Vector Machine (SVM) classifier is employed to vali- date the segmentation results. Proposed character seg- mentation algorithm is evaluated on publicly available Tobacco-800 database and proprietary database.

In the proposed character recognition system, three new features are computed, which are based upon structural geometry of characters. For the first feature, distances between center pixel and character pixels are found out in each non-overlapping block. These non-overlapping blocks are formed with respect to center pixel of char- acter. Later on, row wise and column wise distances are normalized. In second feature, cut vectors are gen- erated in eight different directions.

Bhattacharya and Chaudhari (2005) presented a brief survey on databases for research on recognition of handwritten characters of Indian script. Databases of 22556 samples of Devanagari numerals, 12938 samples of Bangala numerals, 5970 samples of Oriya numerals have been developed. Database of Devanagari numerals is collected from 1049 users. Also 556 users have written Bangala and Oriya numerals.

2. LITERATURE REVIEW:

It is observed that most of the character segmentation and recognition methods reported in the literature are developed for specific script and text (either printed or handwritten). Furthermore, performance of Indic script OCRs is poorer than Latin script OCRs. Hence, the focus of the proposed research work is to segment and recognize the characters from scanned multilingual Indian documents with intermixed texts. The main contributions of the proposed research work are as follows:

For character segmentation, a hybrid scheme is pro- posed where at the beginning; pre-processing is per- formed on word images. Subsequently, one pixel width image, i.e. thinning operation is applied on words and projection profile is used to find primary segmentation paths. Later, distance related criteria are utilized to get fine segmentation paths. During post-processing step, over-segmented and overlapped characters are separated using graph distance theory. Finally, a trained Support Vector Machine (SVM) classifier is employed to validate the segmentation results. Proposed character segmentation algorithm is evaluated on publicly available Tobacco-800 database and proprietary database.

These non-overlapping blocks are formed with respect to center pixel of char- acter. Later on, row wise and column wise distances are normalized. In second feature, cut vectors are gen- erated in eight different directions.

Bhattacharya and Chaudhari (2005) presented a brief survey on databases for research on recognition of handwritten characters of Indian script. Databases of 22556 samples of Devanagari numerals, 12938 samples of These non-overlapping blocks are formed with respect to center pixel of char- acter. Later on, row wise and column wise distances are normalized. In second feature, cut vectors are gen- erated in eight different directions.

Bhattacharya and Chaudhari (2005) presented a brief survey on databases for research on recognition of handwritten characters of Indian script. Databases of 22556 samples of Devanagari numerals, 12938 samples of Bangala numerals, 5970 samples of Oriya numerals have been developed.

CHARACTER RECOGNITION:

Texture features are calculated using transforms like Fourier transform wavelet transform, Gabor transform and Scale

Invariant Feature Trans- form (SIFT). Hartley transform is also utilized in to compute features. Low frequency components from these transforms reflect the basic shape of the character, whereas high frequency components provide detail variations. Conventional transforms like Fourier and wavelet are efficient to capture details in one dimension. In the context of two-dimensional signal like image, these transforms extract details when the two-dimensional sig- nal is represented by collection of one-dimensional signals. In addition, these transforms are unable to handle smooth contours and randomly oriented edges. Gabor filter, some- how overcomes these problems but has spectral limitations. Besides these transforms, gradients features are developed in and whereas run length features are formed.

Structural features are computed by Dash et al. using chords and constellation diagram of characters. Struc- tural stroke features are calculated in and, whereas shape features are proposed . In addition, fitting models and trajectory models are also used to form structural features. The satisfactory perfor- mance from these features describes the importance of stroke and structural information of characters [20]. Conversely, the scope of these methods is very limited up to the characters of one or two scripts because structural features are normally script dependent.

In general, performance of recognition system depends upon the features extracted from characters images. These features should uniquely classify the character in less time. Computing transform feature on the character is a tricky task due to the variation in rotation, direction and frequency. Structural features are very efficient if characters of script (like Indic scripts) are rich in strokes information. If these structural features are normalized, then robustness with respect to font styles, sizes and noise is easily achieved.

PROPOSED WORK

Fig presents the structure of proposed work for character segmentation and recognition. Herein, character segmentation is divided into four phases, whereas character recognition includes training and testing phases. Three features are extracted for recognizing characters.

PROPOSED CHARACTER SEGMENTATION ALGORITHM

The proposed character segmentation algorithm consists of pre-processing, segmentation, post-processing and post-verification using SVM to validate the segmentation results. These steps are described as follows:

PRE-PROCESSING

In pre-processing, gray scale word image I of size X Y having pixel intensity f (m, n) of the pixel located at (m, n) is binarized. Binarization is performed to reduce computational complexity of the algorithm, as only two colors are present for processing. Morphological erosion operation is then per- formed on the binarized word image to join disconnected components as shown in Fig. New novel segmentation algorithm and new structural feature extraction method for recognizing handwritten Marathi compound character. The proposed minutiae detection algorithm gave 95% accuracy for segmentation. By using structural feature edge map getting 94 % recognition accuracy for character recognition.

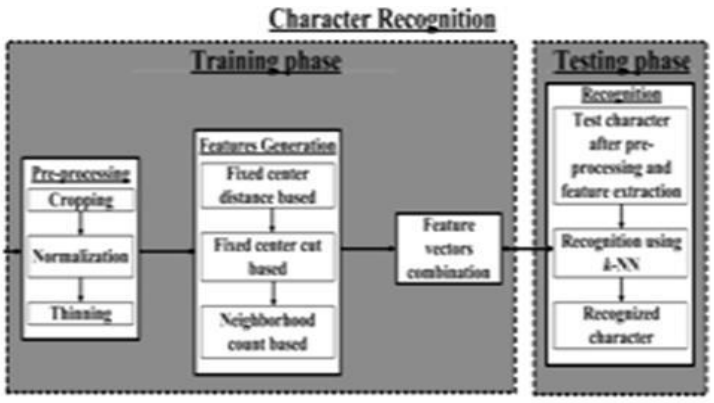

Fig:1 Recognize the character

PROCEDURE FOR THE DETERMINATION OF SAPONIFICATION VALUE

In general, performance of recognition system depends upon the features extracted from characters images. These features should uniquely classify the character in less time. Computing transform feature on the character is a tricky task due to the variation in rotation, direction and frequency. Structural features are very efficient if characters of script (like Indic scripts) are rich in strokes information. If these structural features are normalized, then robustness with respect to font styles, sizes and noise is easily achieved.

3. RESULT AND DISCUSSION:

By using OCR and KNN we recognized the characters in historical documents and Overlapped and over- segmented characters are also separated accurately during post-processing and post-verification stages, respectively. For character recognition, three new structural geometry based features are proposed. FCDF and FCCF are calculated with respect to center pixel of thinned character image, whereas NCF is calculated using neighborhood information of text pixels.
The paper presents a system for handwritten devnagri character recognition for Devnagri script. A huge character dataset is collected from various writers and used for database creation for neural network training and testing. The recognition of characters is done using multistage multi-feature hybrid recognition scheme.

4. CONCLUSION:

In this research, we focused on investigating use of suitable preprocessing and features for effective HCR for Indian Languages, with devnagari being used as the base. The work started with building an extensive corpus for the experiment covering basic characters and Kagunita. The major focus was on preprocessing and feature identification and extraction. The preprocessing stage posed major challenge as the samples of the corpus used for the experiment showed huge variations resulting in significant deformation in the processed image. We initially built a static preprocessing pipeline and based on the noise analysis of resulting images, we proposed a dynamic preprocessing pipeline that mostly eliminated the noise

REFERENCES:

1. Feature Extraction Techniques Implementation Review and Case Study Uma Bhati Department of Computer Science & Engineering JSS Academy of Technical Education Noida-201301
2. A Review of Research on Devnagari Character Recognition Vikas J Dongre Vijay H Mankar Department of Electronics & Telecommunication, Government Polytechnic, Nagpur, India
3. Segmentation of Marathi Handwritten Characters and Numerals Ratnashil N Khobragade Assistant Professor, P G Dept of CS, SGB Amravati University, Amravati, Maharastra, India
4. A Streamlined OCR System for Handwritten Marathi Text Document Classification and Recognition Using SVM-ACS Algorithm Surendra Pandurang Ramteke Department of Electronics & Telecommunication Engineering, Shram Sadhana Bombay Trust College of Engineering and Technology, Bambhori, Maharashtra, India
5. Feature Extraction for Marathi Compound Character Using Edge Map Mrs.Snehal S.Golait Research Scholar ,Department of Computer Science and Engineering, G.H.Raisoni College of Engineering,Nagpur,
6. Zernike Moment Feature Extraction for Handwritten Devanagari (Marathi) Compound Character Recognition Karbhari V. Kale, Department of Computer Science and IT, Dr. B. A. M. University, Aurangabad, Maharashtra, India – 43
7. Maliki, M. & Ifijen, I. H. (2020). Extraction and Characterization of Rubber Seed Oil. International Journal ofScientific Engineering and Science, 4(6), 24-27.
8. Y. Chen and G. Leedham, “Decompose algorithm for thresholding degraded historical document images,” IEE Proc. Vis., Image Signal Process., vol. 152, no. 6, pp. 702–714, Dec.
9. J. G. Kuk, N. I. Cho, and K. M. Lee, “Map-MRF approach for binarization of degraded document image,” in Proc. Int. Conf. Image Process., 2008, pp. 2612–2615.
10. S. Kumar, R. Gupta, N. Khanna, S. Chaudhury, and S. D. Joshi,“Text extraction and document image segmentation using matched wavelets and MRF model,” IEEE Trans Image Process., vol. 16, no. 8, pp. 2117–2128, Aug. 2007.
11. Handwritten Devanagari Compound Character Recognition Juhee Sachdeva, Ph.D. Scholar, Jaipur National University, Jaipur,India,