Original Article
ISSN (Online): 2582-7472

ShodhKosh: Journal of Visual and Performing Arts
June 2024 5(6), 9–15

EXPLORING FEATURE EXTRACTION TECHNIQUES FOR CHARACTER RECOGNITION IN HANDWRITTEN MODI SCRIPT DOCUMENTS

Ketki R. Ingole [1], Dr. Pritish Tijare [2]

[1] Research Scholars, Department of Computer Science and Engineering Sipna College of Engineering and Technology Amravati, India
[2] Professor, Department of Computer Science and Engineering Sipna College of Engineering and Technology Amravati, India

ABSTRACT

From the 19th century Devangari scripting is used for documentation in Maharashtra. Before Devanagari, Modi script were used mostly for documentation purpose, but due to cursive writing style, complex nature, and lack of awareness it diminishes over time. Modi script has historical roots and large number of manuscripts available in various Museums,National libraries and at cultural heritages. With time the knowledge within these documents remains unrevealed, due to lack of proper sources. The Government takes steps to preserve these documents and uncover the information within them. Now a days, emerging technologies pave the way to restore cultural heritage by using HOCR.In this study, we focus on recognizing script data from image documents, particularly handwritten Modi script. This paper explores various feature extraction techniques that facilitate character recognition from these historical handwritten documents.

Keywords: Devanagari, Modi-Script, HOCR, Manuscripts

Corresponding Author

Ketki R. Ingole, mohodketki@gmail.com

DOI
10.29121/shodhkosh.v5.i6.2024.1631

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

1. INTRODUCTION

One key area of focus for field researchers is Optical Character Recognition (OCR), a branch of Computer Vision and Pattern Recognition. While significant progress has been made in recognizing printed text, one of the most challenging aspects remains the recognition of handwritten text. Handwritten OCR poses a unique challenge due to the variability in individual handwriting styles. The difficulty is further amplified when the task involves recognizing characters from ancient manuscripts, as these scripts often present additional complexities. Advancing research in the recognition of ancient scripts is crucial for uncovering the knowledge preserved in historical manuscripts.

Extensive research has been conducted on foreign languages, yielding positive results in Optical Character Recognition (OCR). In India, however, recognizing text presents unique challenges due to regional variations in languages, where spoken and written forms can differ significantly. Indian scripts, with their complex nature, further complicate this task. Many scripts have been influenced by historical kingdoms, social environments, and regional languages. According to a survey, an article [1] highlights that cursive writing is a significant hurdle in Intelligent Character Recognition (ICR), as it is difficult to segment and recognize characters that are written in continuous patterns.

In Maharashtra, the regional script Modi, which evolved over time influenced by various ruling dynasties, was eventually replaced by the Devanagari script by the late 20th century. Modi script, used for writing Marathi, has a rich history spanning 700 years. It originated in the 12th century during the Yadav Dynasty, saw development during the Bahamani era, and was further refined in the Chhatrapati Shivaji and Peshwa eras. After India's independence, the Devanagari script gradually replaced Modi. Modi was also used in Tamil and Gujarati. This paper aims to identify and describe the fundamental structural features of Modi script letters, focusing on the components and characteristics that distinguish one letter from another, which is essential for effective feature extraction.

Modi script lacks punctuation marks or delimiters to indicate the beginning and end of sentences and words, which complicates reading. It comprises 46 letters: 36 consonants and 10 vowels, whereas Devanagari script includes 48 letters: 36 consonants and 12 vowels. The fewer vowels in Modi script help reduce grammatical errors compared to Devanagari. Modi script is written by drawing a horizontal line across the page from left to right, known as 'Shirolekh', with letters positioned according to this line.

Historically, Modi script was widely used for official and administrative purposes, as well as for religious texts, literature, and correspondence. It was especially common during the Maratha Empire and remained in use until the 20th century. The introduction of the British Raj and the spread of printing technology led to a decline in the use of Modi script, favoring Devanagari and other scripts.

2. LITERATURE REVIEW

Researchers have explored the recognition of various Indic and medieval scripts, including the Modi script. This script, with its historical and literary significance in central and western India, has evolved over time, resulting in significant variations. Despite its importance, automating Optical Character Recognition (OCR) for the Modi script remains relatively underexplored. Initial research by D. N. Besekar et al. [2], [3] has focused on recognizing the Modi script. Paper [2] examines the structural similarities between standard and handwritten Modi characters, while paper [3] provides a theoretical analysis of the script and the challenges in its recognition. Their specific efforts include mathematical morphological approaches [4] and chain code-based recognition [6], though chain code accuracy may vary due to the script's free-form writing style and the similarity of many characters.

Sidra Anam and colleagues [7] developed the Modi Script Character Recognizer System (MSCR), utilizing Otsu's Binarization and the Kohonen neural network approach. This system, trained on 22 distinct Modi characters from handwritten samples, demonstrated effectiveness but had lower recognition rates for similarly shaped characters, achieving a 72.6 percent accuracy for handwritten characters.

Ramteke A.S. and Katkar G.S. [8] used a structural similarity method to identify Modi characteristics, achieving a recognition rate of 91 to 97 percent with Measured Structure Similarity (SSIM), KNN, and Backpropagation Neural Networks. Besekar D.N. and Ramteke R.J. [9] applied a zone-based technique for offline handwritten digit recognition, achieving a 93.5 percent accuracy with a variance table after preprocessing steps like size normalization and noise removal.

Theoretical analysis of Modi script recognition by Besekar D. N. & Ramteke R. J. [10] highlighted the challenges in extracting structural elements due to the script's cursive form, character variations, and handwriting styles. The study recommended improvements for segmentation and recognition.

To enhance Modi character recognition, [11] proposed using a CNN autoencoder for feature extraction, reducing the feature set size from 3600 to 300, followed by SVM classification, which achieved a 99.3% accuracy, surpassing other methods.

Savitri Chandure and Vandana Inamdar [12] developed a supervised Transfer Learning (TL) based classification system, constructing a dataset for Modi handwritten characters. They utilized a pre-trained Deep Convolutional Neural Network (DCNN) AlexNet for feature extraction and SVM for classification, achieving recognition accuracy rates of 92.32% for Modi characters and 97.25% for Devanagari characters.

3. STRUCTURAL FEATURES OF MODI SCRIPT

A structural approach to character recognition focuses on analyzing the fundamental components and organization of characters, rather than relying solely on pixel intensity values or low-level image features. This method examines the hierarchical arrangement, spatial relationships, and geometric properties of the character elements.

For recognizing Modi characters, a structural approach would involve analyzing features such as strokes, curves, intersections, and other defining structural elements of the script. Techniques used in this approach might include:

- **Contour Analysis:** Examining the outlines of characters to identify their shapes and boundaries.
- **Stroke Decomposition:** Breaking down characters into their individual strokes to understand their formation.
- **Graph-Based Representation:** Using graph structures to represent the relationships and connections between different parts of a character.

By focusing on these structural aspects, the recognition system can better accommodate variations in handwriting styles, document degradation, and other challenges associated with historical manuscripts. This approach enhances the system's ability to accurately recognize characters despite inconsistencies and distortions.

Figure 1

Figure 1 Consonants in Modi Scripts

Figure 2

Figure 2 Vowels in Modi Script

4. FEATURE EXTRACTION TECHNIQUES

For feature extraction from ancient Modi manuscripts, the best method may depend on various factors such as the condition of the manuscripts, the style of handwriting, and the specific characteristics of the Modi script. However, considering the nature of ancient manuscripts, pre-processing is required by enhancing the image quality and reducing the noise present in the image document and segmentation. Despite this challenges associated with their analysis, some feature extraction methods that might be suitable include:

1) Convolutional Neural Network (CNN):

As per the analysis Convolutional Neural Network performs excel at image-based tasks such as character recognition by automatically learning hierarchical features from input images. This characteristics of CNN might be give effective results for complex Modi script, where characters have intricate details and diacritical marks.The input is a binary image of the character, represented as a matrix of pixels value.

The Convolutional operation is defined as:

$$(X * K)_{i,j} = \sum_{n}^{n} X_{i+m-1,j+n-1} \cdot K_{m,n}$$

This generates the feature map highlighting edges or other features in the image. To learn complex pattern og handwritten text apply a non-linear activation function like ReLu (Rectified Linear Unit) to the feature map:

$$f(x)=\max(0, x)$$

Later on pooling layer helps to reduce the size of the feature map. Use Softmax function at the final output layer to predict the probability of the specific characters.

$$\text{Softmax}(z_i) = \frac{e^{z_i}}{\sum_j e^{z_j}}$$

2) Convolutional Neural Network (CNN)+Histogram of Oriented Gradients (HOG):

The Modi script generally has different structural characteristics for each character at the same time manuscript might be facing the low resolution and noise problem. Hence to enhance the model performance the CNN can be combined with the traditional feature extraction technique.

HOG is the excellent for capturing edge and gradient information which is important for distinguishing between characters that are in similar shape but different in finer details.

At first compute gradients Gx and Gy along the x and y directions of the image.

$$Gx = \frac{\partial X}{\partial x}, Gy = \frac{\partial X}{\partial y}$$

And calculate the magnitude and direction of the gradient at each pixel and create histograms of the gradient directions within the local regions of the image.

$$M = \sqrt{G_x^2 + G_y^2}$$

The CNN processes the HOG features through Convolutional and fully connected layers to classify the character. HOG captures the edge orientations, distinguishing characters with similar shapes but different strokes. Creating a hybrid approach of HOG+CNN allows the model to focus detailed texture information.

3) Support vector Machine (SVM) with Principal Component Analysis (PCA):

Principal Component Analysis is one of the powerful tool for dimensionally reduction, help in simplifying the feature space and improving the efficiency of recognition models. It reduces the dimensional of the feature set obtained from traditional feature extraction methods before feeding it into the classifier like an SVM.

Support vector Machine (SVM) are effective for classification tasks, when feature space is well defined, it work with traditional feature extracion methods for charcter recognition. After extracting features, SVM is used to classify the Modi Characters. It performs well with smaller datasets.

Suppose the image has n pixels. Then flatten the image into a vector and compute the covariance matrix $\sum$ of the dataset.

$$\sum = \frac{1}{m} \sum_{i=1}^{m} (x_i - \mu)(x_i - \mu)^T$$

Then perform eigen decomposition to get eigenvalues and eigen vectors,

$$\sum v = \lambda v$$

Then select the top k eigenvectors corresponding to the largest eigenvalues to form a matrix Vk

Support vector Machine (SVM) use the feature vector y as input to an SVM. The SVM finds the optimal hyperplane that separate the classes.

4) Zoning:

Zoning divides the character into smaller zones and extracts statistical features like pixel density within each zone. This method helps capture local variations within a character.

Divide the character image into predefined zones and extract features such as number of pixels in each zone. These can then be classified using traditional classifiers or neural networks.

To recognize the Modi character, divide the character image into n X n zones and calculate the pixel density in each zone.

$$z_{i,j} = \frac{\text{Number of black pixels in zone}(i, j)}{\text{Total pixels in zone}(i, j)}$$

The feature vector Z is composed of the densities of all zones. Later feature vector Z can be fed into simple classifiers or a neural network.

Zoning captures the distribution of pixels within specific regions of the character represents the shape effectively. The feature vector is low dimensional representation of the character, which used in distinguishing similar charcters.

5. CONCLUSION

The recognition of manuscript is more complicated as compare to printed documents. The main hurdles for implementation of automated recognition system for Modi script are the complexity of the script, cursive writing, overlapping of characters, oldest scripting language, different handwriting style. Another major hurdle is cursive writing style, without punctuation marks, structural similarities between characters, conjunction of consonants and vowels.

The study of structural behaviour of characters of Modi Script helps for extraction of features, finding appropriate feature extraction approach. At the same time analyzing various feature extraction technique helps to design proper model for recognizing the Modi characters from manuscripts.

It is identified that the CNN are powerful for character recognition due to their ability to learn the complex features. CNN with HOG can enhance the performance of the model. SVM with PCA is useful when computational resources are limited and Zoning is simple effective for capturing the overall structure of character.

These methods can significantly improve the accuracy of Modi character recognition systems.

CONFLICT OF INTERESTS

None.

ACKNOWLEDGMENTS

None.

REFERENCES

A S Ramteke, G S Katkar, Recognition of Off-line Modi Script : A Structure Similarity Approach, “International Journal of ICT and Management” ISSN No. 2026-6839, February 2013

D. N. Besekar, A S Ramteke, “Theoretical analysis of MODI script according to recognition point of view, some issues involved with character recognition of MODI script”, International Journal of Computer Applications, February 2013

D. N. Besekar, “Recognition Of Numerals Of Modi Script Using Morphological Approach”,Shodh Samiksha Aur Mulyankan Vol.III, Issue-27, april 2011

Anil K. Jain, Template-based online character recognition, Pattern Recognition, Volume 34, Issue 1, January 2001, Pages 1–14

D. N. Besekar, A S Ramteke, “Chain Code Approach For Recognizing Modi Numerals , Indian Journal Of Applied Research”, December 2011.

Algorithm and Kohenen Neural Network," International Journal of Computer Applications (0975 –8887) Volume 111 – No 2, February 2015.

Ramteke A.S., Katkar G.S., 2012, "Recognition of Offline MODI Script," International Journal of Research in Engineering, IT and Social

Besekar D.N., Ramteke R.J., 2012, "Feature Extraction Algorithm for Handwritten Numerals Recognition of MODI Script using Zoning-based Approach," International Journal of Systems, Algorithms & Applications, Volume 2, Issue ICRASE12, ISSN 2277 2677, pp. 1-4.

Besekar D.N., Ramteke R.J, 2013, “Study for Theoretical Analysis of Handwritten MODI Script – A Recognition Perspective,” International Journal of Computer Applications, vol. 64, no. 3, ISSN 0975-8887, pp. 45-49.

S. Joseph and J. George, "Handwritten Character Recognition of MODI Script using Convolutional Neural Network Based Feature Extraction Method and Support Vector Machine Classifier," 2020 IEEE 5th International Conference on Signal and Image Processing (ICSIP), 2020, pp. 32-36.

S. Chandure and V. Inamdar, "Handwritten Modi character recognition using transfer learning with discriminant feature analysis," IETE Journal of Research, pp. 1–11, 2021.

Manisha s. Deshmukh, Manoj Patil, and Satish Kolhe," Offline handwritten Modi numerals recognition using chain code." WCI 15.

Ankit Kumar Sahכ, Showmik BhowmikŞ, Samir Malakarş, Ram SarkarŞErgina KavallieratouŢ, Nikos Vasilopoulos “Text and Non-text Recognition using modified HOG descriptor”,IEEE Xplore,05 Feb 2018