Recognition of Marathi Characters using PCA Algorithm

Shubham Arun Fate
PCE Nagpur

Nikhilesh R Lingayat
PCE Nagpur

Snehal Golait
PCE Nagpur

ABSTRACT
The developing need have written by hand Marathi character acknowledgment in Indian workplaces, for example, international ID, railroads and so on has made it key range of an examination. Because of expansive character set, Complex shape are more inclined to miss characterization. Highlight extraction is one of the fundamental capacity of manually written Script Identification. It includes measuring those components of the data example are important to order. This paper proposed a PCA calculation to perceive transcribed Marathi character PCA is a method for recognizing designs in information and communicating the information so as to highlight their similitudes and contrasts. Chief part examination (PCA) is a traditional measurable technique. This straight change has been broadly utilized as a part of information investigation and pressure. Main segment examination depends on the factual representation of an irregular variable. The PCA system views every character picture as a component vector in a high dimensional space by linking the columns of the picture and utilizing the power of every pixel as a solitary element vector.

Key Words
Segmentation, PCA, Binarization

1. INTRODUCTION
Optical Character Recognition (OCR) is a procedure of programmed acknowledgment of various characters from an archive picture. OCR frameworks are considered as a branch of counterfeit consciousness and a branch of PC vision too. Scientists arrange OCR issue into two spaces. One manages the picture of the character by filtering which is canceled line acknowledgment. Alternate has distinctive data way, where the essayist composes specifically to the framework utilizing, for instance, light pen as a device of information. This is called online recognition. Fig (1) shows the block diagram of the typical OCR system.

Conventional OCR frameworks are experiencing two primary issues, one originates from highlight extraction stage and alternate originates from classifier. Highlight extraction stage is in charge of extricating components from the picture and passing them as worldwide or neighborhood data to the following stage with a specific end goal to help the later taking choice and perceiving the character.

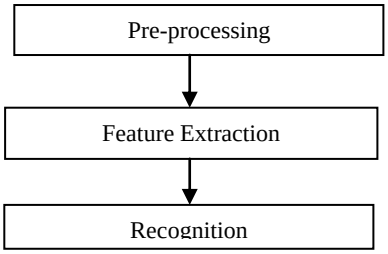


Figure 1: Typical OCR block diagram

Two difficulties are confronted: if highlight extractor extricates numerous elements keeping in mind the end goal to sufficiently offer data for classifier, this implies numerous calculations and in addition more intricate calculations are required. In this way it is time expended process. On the other hand, if couple of segments are evacuated with a particular deciding objective to quicken the system, lacking information may be gone to classifier. The second essential issue that classifier is responsible for, is that most of classifiers rely on upon Artificial Neural Networks (ANNs). Be that as it may, to enhance the knowledge of these ANNs, colossal emphases, complex calculations and learning calculations are required, which likewise prompt expend the processor time. In this manner, if the acknowledgment exactness is enhanced, the expended the truth will surface eventually and the other way around.

To handle these issues, another OCR development is not proposed in this paper, where highlights extractor nor is ANN required. The proposed improvement relies on upon the photo weight procedure by then delivers a novel vector code identifying with the entire picture. At that point produces a novel vector (code) relating to the whole picture. This vector can be viably used to perceive the character since it conveys the fundamental subtle elements of the character's picture. The criticalness of the rule subtle parts is that they are standard among the same character which is framed by various essayists. This vector can be viably used to perceive the character since it conveys the fundamental subtle elements of the character's picture

2. LITERATURE REVIEW
Manually written character acknowledgment is the vital range in picture preparing and design acknowledgment fields. Acknowledgment in Indian script is a testing assignment exceptionally Devanagari. Chip away at Devanagari was begun right on time in 1970. Initially investigate report on transcribed Devanagari character was distributed in 1977 by Sethi and Chatterjee.

A broad examination on printed Devanagari content was completed by Veena Bansal and R.M. K. Sinha [8, 9]. To begin with framework for manually written numeral acknowledgment of Devanagari characters was proposed by R. Bajaj, P.M. Patil and T .R. Sontak likewise displayed a calculation for manually written Devanagari numeral acknowledgment which was turn, scale and interpretation invariant. U. Buddy et al. [11] displayed a framework for disconnected from the net written by hand character acknowledgment of Devanagari utilizing directional data for separating highlights. Multilayer perceptron was additionally utilized for order by Sandhya Arora et al. for written by hand Devanagari characters alongside various components independently. The last arrangement result was acquired by a choice calculation in light of weighted larger part voting procedure. In U. Buddy et al consolidated two classifiers to get higher exactness of Devanagari character acknowledgment with the same elements. Joined use of SVM and Modified Quadratic Discriminant Function (MQDF) are associated for better execution of Devanagari character affirmation. Starting late, a relative examination of various components and classifiers used for composed by hand Devanagari character

affirmation was done by U. Pal et al. [11]. They found that Mirror Image classifier was the best classifier. Affirmation of physically composed Bangla compound character was attempted by U. Mate et al. [11] using slant highlights.

3. PRE-PROCESSING

The fundamental favorable position of preprocessing a transcribed character picture is to sort out the data in order to make the errand of acknowledgment less difficult. The preprocessing incorporates the binarization, standardization, skew redress and division.

3.1 Binarization

This is the initial phase in the preparing of examined picture. In this procedure first digitization of Image is finished. For binarization thresholding procedure is utilized. Thresholding is a picture handling method for changing over a dim scale or shading picture to a double picture based upon an edge esteem. In the event that a pixel in the picture has a power esteem not exactly the limit esteem, the comparing pixel in the resultant picture is set to dark. Something else, if the pixel power worth is more prominent than or equivalent to the edge force, the subsequent pixel is set to white. In this way, we have utilized a picture with just 2 hues, dark (0) and white (255).

3.2 Noise reduction

Amid the examining process, some twisting in pictures might be acquainted due with low quality of pen, light hand penmanship, and poor paper quality on which the characters are composed and so on.

3.3 Normalization

Standardization is one of the imperative pre-handling elements for character acknowledgment. Ordinarily, in normalization the character picture is directly mapped on to a standard plane by interjection/extrapolation. The size and position of character is controlled such that the length and width of normalized p lane are filled. By linear mapping, the character shape is not only deformed but also the aspect ratio changes.

3.4 Skew Correction

Skew Correction methods are used to align the paper document with the coordinate system of the scanner.

3.5 Segmentation

The pre-preparing stage gives a spotless record in which least clamor picture is acquired. The following stage is portioning the report into its sub parts and extricating the significant elements to nourish the preparation and acknowledgment stages.

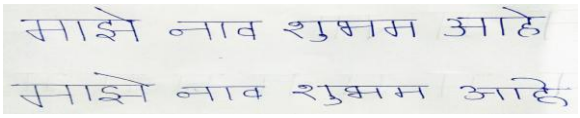
There are three stages in division as takes after:-

- Line Segmentation
- Word Segmentation
- Character Segmentation

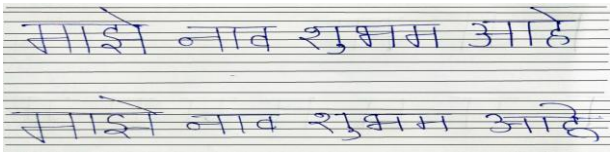
Line Segmentation

In Line Segmentation it is going to project horizontal projection i.e. row-wise projection to separate the lines.

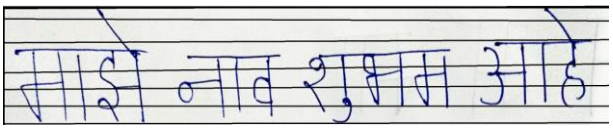
Example: - Sample



Horizontal projection



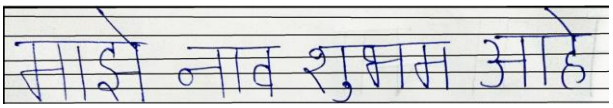
Line is segmented



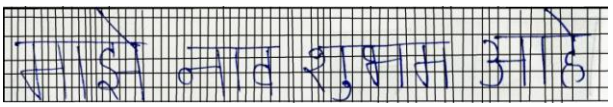
Word Segmentation

In Word Segmentation it is going to project vertical projection on segmented line.

Example: Segmented Line Sample



Vertical Projection



Word are Segmented

Character Segmentation

After Word Segmentation the Character is segmented through scanning in horizontal way by marking starting point and ending point of a character.

Example: Word Segment Sample

Character is segmented

4. FEATURE EXTRACTION

The real objective of highlight extraction is to separate an arrangement of components. It boosts the acknowledgment rate with minimal measure of components.

They are extensively grouped into two distinct classifications.

- Statistical features
- Structural features

To perceive written by hand Marathi character we are utilizing PCA calculation for highlight extraction.

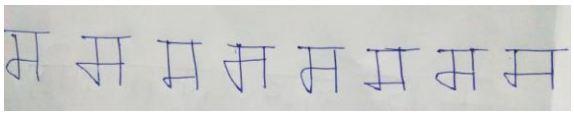
PCA (Principal Component Analysis)

Foremost segment examination depends on the factual representation of an irregular variable. The proposed calculation for PCA is as per the following.

- Step 1: Get a few information

May there be N Characters (A1, A2, - , A) constituting the preparation set meant by m by n networks and indicated as framework A.

Example:



- Step 2: Subtract the mean:

The PCA to work well, you need to figure mean from highlight vector. After this mean is subtracted from the first grid. Assume an is unique framework and Ā is mean, the outcome is put away in Φi. Recipe for

$$\bar{A} = \frac{1}{N} \sum_{i=1}^{N} A_i$$

$$\Phi i = Ai - \bar{A}$$

- Step 3: Calculate the covariance grid

The Covariance grid C is computed by utilizing the equation

$$C = \frac{1}{N} \sum_{i=1}^{N} \Phi_i * \Phi_i^T$$

- Step 4: Calculate the eigenvectors and Eigen estimations of the covariance grid
  Eigenvalue is figured from the equation
  $$(A - \lambda I) = 0$$
  Where, λ is eigenvalue and I is identity matrix.

- Step 5: Choosing parts and shaping a component vector

  Pick the eigenvector with the most elevated eigenvalue is the guideline segment of the information set

- Step 6: Deriving the new information set

Final Data = Row feature vector * Row Data Adjust

5. CONCLUSION

There are various methods of feature extraction in that depending on the features the technique for extracting the features are developed and then depending on that classification of features is done. PCA is one of the technique for feature extraction. This method to extract the features of handwritten Marathi character.

6. REFERENCES

[1] Ms.Snehal Dalal, Dr.Latesh Malik, "HANDWRITTEN SCRIPT IDENTIFICATION FOR INDIAN POSTAL AUTOMATION WITH PRINCIPAL COMPONENT ANALYSIS**" In**ternational Conference "MNGSA-08" at Coimbatore.

[2] Prof. M. S. Kumbhar and Y. Y. Chandrachud "HANDWRITTEN MARATHI CHARACTER RECOGNITION USING NEURAL NETWORK" International Journal of Emerging Technology and Advanced Engineering Website: www.ijetae.com (ISSN 2250-2459, Volume 2, Issue 9, September 2012).

[3] R. Plamondon and *S.* N. Srihari, "On-line and off-line handwritten recognition: .A comprehensive survey", IEEE Trans. on Patten Analysis and Machine Intelligence, Vol. *22*, 2000, pp. 62-84.

[4] U. Mahadevan, and S. N. Srihari, "Parsing and Recognition of City, State, and ZIP Codes in Handwritten Addresses", In Proc. of 5th Int. Conf. on Document Analysis and Recognition, 1999,

[5] R. Seiler M. Schenkel E Eggimann Swiss Federal Institute of Technology, Zurich [Off-Line Cursive Handwriting Recognition Compared with On-Line Recognition]

[6] G. Boccignone, A. Chianese, L. Cordella, and A. Marcelli. Recovering dynamic information from static handwriting. Patten Recognition, 26(3):409-418, 1993.

[7] U.Mahadevan, and S. N. Srihari, "Parsing and Recognition of City, State, and ZIP Codes in Handwritten Addresses", In Proc. of 5th Int. Conf. on Document Analysis and Recognition, 1999, pp.

[8] Veena Bansal, and R. M. K. Sinha, "On How to Describe Shapes of Devanagari Characters and Use them for Recognition", Proc. 5th Int. Conf. Document Analysis and Recognition, Bangalore, India, Sept. 20-22, 1999, pp. 410-413.

[9] Veena Bansal, "Integrating Knowledge Sources in Devanagari Text Recognition", Ph.D. Thesis, IIT Kharagpur, India, 1999.

[10] K. Roy, S. Vajda, U. Pal, and B. B. Chaudhuri, "A System towards Indian Postal Automation". In Proc. of international Workshop on Frontier of Handwriting Reognition-9, 2004.

[11] U. Pal and B. B. Chaudhuri, "Script line separation from Indian multi-script documents" IETE Journal of Research, Vol. 49, 325-328. 2003, pp. 3-1 1.

[12] U. Pal and P. P. Roy."Multi-oriented and curved text lines extraction from Indian documents", IEEE Trans. on Systems, Man and Cybernetics- Part B, Vo1.34, 2004, pp. 1676-1684.