IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

Impact Factor: 7.301

Volume 3, Issue 16, May 2023

Sanskrit Text Recognition using Machine Learning

Pankaj Tukaram Bhise, Vina Lomte, Darshan Appa Derle, Pratik Bhagwat Gitte, Onkar Vilas Lonari

Department of Computer Engineering

RMD Sinhgad School of Engineering, Pune, India

Abstract: *Character recognitions are becoming increasingly important as technology continues to improve at an astounding rate, and they serve a vital role in encouraging research into OCR techniques. Researchers have discovered that the identification of Sanskrit handwriting is one of the most difficult areas of research in the field of pattern recognition. Using character recognition software, you may encode handwritten or printed text from scanned photographs. The software turns the data into a format that can be read by machines. When it comes to the verification of individuals and documents, character recognition is a biometric capability that is commonly employed. An off-line handwritten character recognition system was developed in this research using a feed forward neural network as the input to the network. Using a handwritten Sanskrit character sized to 20x30 pixels, the neural network is trained to recognise words in English. Following the training process, neural networks with different sets of hidden neurons were trained and their identification rates for Sanskrit characters were compared against each other. According to the results of the proposed system, the accuracy rates are comparable to those of earlier handwritten character recognition systems in terms of speed and accuracy.*

I. INTRODUCTION

Science, mathematics, Hindu mythology, Indian civilization, and culture can all be found in ancient Sanskrit writings. For the sake of sharing this knowledge with the world and encouraging further research into this ancient literature, it is imperative that these texts are easily accessible. Our work here presents a Convolutional Neural Network (CNN)-based Optical Character Recognition (OCR) system that can accurately digitise ancient Sanskrit manuscripts (Devanagari Script) despite their poor state.

An image segmentation technique that estimates pixel intensities is used to identify the letters in the image in order to identify them. In order to improve the accuracy of segmentation, the OCR separates common compound characters (half-letter combinations) into their own classification. Sanskrit manuscripts that are unclean and poorly maintained make excellent candidates for optical character recognition (OCR) because of the technology's resistance to changes in image quality, picture contrast, typeface, and letter size.

II. LITERATURE SURVEY

Sr. No.	Author Name/ Paper Name	Proposed System	Advantages	Disadvantages
1	Mande Shen, Hansheng Lei, “Improving OCR Performance with Background Image Elimination”	They present a novel and cost-effective image pre-processing method to accomplish the task. They first enhance the document images before OCR by utilizing the brightness and chromaticity as contrast parameters. Then they convert color images to gray and threshold it. This way, background images can be removed effectively without losing the quality of text characters.	Effective in removing thebackground image and thus enhances the performance of OCR.	Problem of black-white backgrounds.

DOI: 10.48175/568

2	Vivek Shrivastava and Navdeep Sharma, "ARTIFICIAL NEURAL NETWORK BASED OPTICAL CHARACTER RECOGNITION"	Certain topological and geometrical features are determined in order to classify and recognise a character accurately. Overall shape and features like lines, curves, protrusions etc. are also interpreted by the human mind. Spatial pixel-based calculation is used to extract these properties, also known as Features, from the image.	Because there are just a small number of features to calculate, the system is less time-consuming, has a tiny database, and is highly adaptable to untrained inputs.	Unable it to recognize stylized fonts also
3	Rokus Arnold, Poth Miklos, "Character Recognition Using Neural Networks'	Sample recognition is a typical problem for neural networks to handle. Character recognition is one of these. This is one of the most straightforward neural network implementations. Using Matlab's Neural Network Toolbox, they sought to distinguish printed and handwritten characters by projecting them into different sized grids (5x7, 7x11, and 9x13).	Less Time Complexity, Good Extraction Result	Problem of black-white backgrounds.
4	Vedgupt Saraf, D.S. Rao, "Devnagari Script Character Recognition Using Genetic Algorithm for Get Better Efficiency	For this reason, a genetic algorithm is a great way to combine several types of character writing and create new ones. Despite the fact that they may be viewing a writing style for the first time, scientists have observed that individuals are nevertheless able to distinguish characters in handwriting that they have never seen before.	Proposed system obtained 98.78% recognition accuracy.	Unable it to recognize stylized fonts also
5	P.Murugeswari, Dr.D.Manimegalai, "Complex Background and Foreground Extraction in Color Document Images using Interval Type-2 Fuzzy"	For the processing of colour document images, a new interval type-2 fuzzy thresholding method is proposed in this paper. Experiments with the proposed method have used a variety of different backdrop textures and colour schemes as well as different types of foregrounds and background text.	There is no need to manually fine-tune the proposed approach because it uses automatic parameter estimation. It can also handle papers with a variety of complex backgrounds.	Not work on degraded document binarization using the Type-2 fuzzy.

III. PROPOSED SYSTEM

1.Input Image: Here we can upload the Input Image.

2.Image Pre-processing:

In this step we will apply the image pre-processing methods like grey scale conversion, image noise removal.

3.Image segmentation: in this step we will apply the segmentation using K-means clustering.

4.Image Feature Extraction:

In this step we will apply the image pixel extraction methods to remove the image features from image.

5.Image Classification:

In this stage we will apply the picture classification methods to distinguish features.

6.Result:

In this step will show the final result.

IV. ALGORITHMS

A.K-Means Clustering

K-Means is the one of the unsupervised learning algorithms for clusters. Clustering the image is grouping the pixels according to the same characteristics. In the k-means algorithm initially we have to define the number of clusters k. Then k-cluster centerare chosen randomly. The distance between each pixel to each cluster centers are calculated. The distance may be of simple Euclidean function. Single pixel is compared to all cluster centers using the distance formula. The pixel is moved to particular cluster which has shortest distance among all. Then the centroid is re-estimated. Again, each pixel is compared to all centroids. The process continuous until the center converges.

Flowchart of K-Means Algorithm

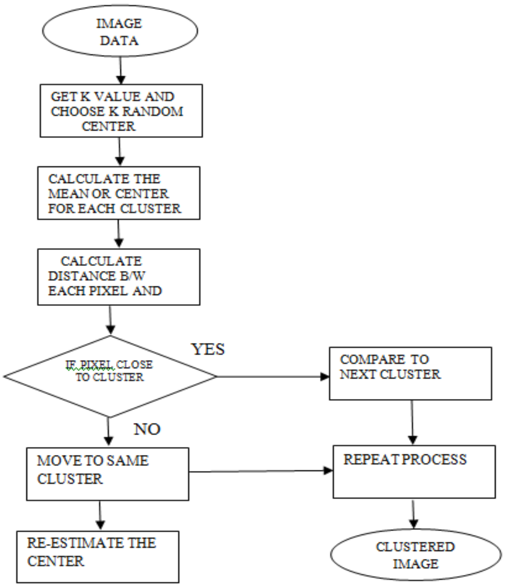


IMAGE DATA
GET K VALUE AND CHOOSE K RANDOM CENTER
CALCULATE THE MEAN OR CENTER FOR EACH CLUSTER
CALCULATE DISTANCE B/W EACH PIXEL AND
IF PIXEL CLOSE TO CLUSTER
YES
COMPARE TO NEXT CLUSTER
NO
MOVE TO SAME CLUSTER
REPEAT PROCESS
RE-ESTIMATE THE CENTER
CLUSTERED IMAGE

Figure 4: K-Means Flow Chart

B CONVOLUTION NEURAL NETWORK

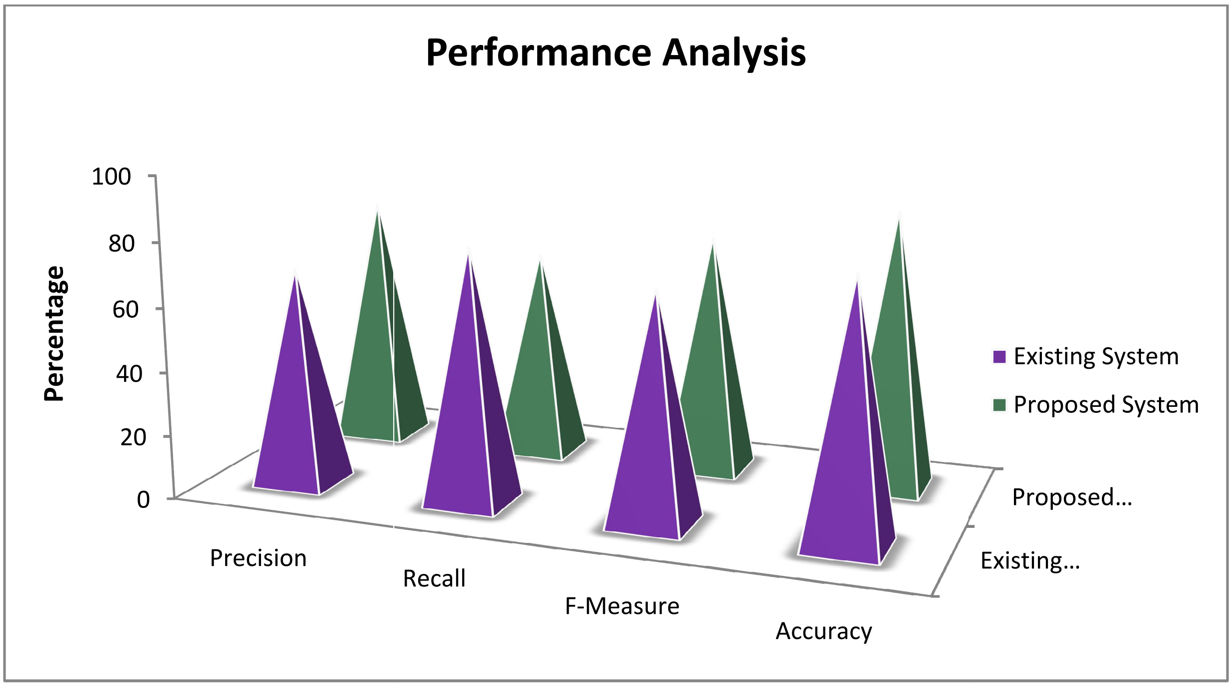
The structure of CNN algorithm includes two layers.First is the extraction layer of features in which each neuron's input is directly connected to its previous layer's local ready fields and local features are extracted. The spatial relationship between it and other features will be shown once those local features are extracted. The other layer is feature_map layer; Every feature map in this layer is a plane, the weight of the neurons in one plane are same. The feature plan"s structure make use of the function called sigmoid. This function known as activation function of the CNN, which makes the feature map have shift in difference. In the CNN each convolution layer is come after a computing layer and its usage is to find the local average as well as the second extract; this extraction of two feature is unique structure which decreases the resolution.

Step 1: Select the dataset.
Step 2: Perform feature selection using information gain and ranking
Step 3: Apply Classification algorithm CNN
Step 4: Calculate each Feature fx value of input layer
Step 5: Calculate bias class of each feature
Step 6: The feature map is produced and it goes to forward pass input layer
Step 7: Calculate the convolution cores in a feature pattern

Step 8: Produce sub sample layer and feature value.
Step 9: Input deviation of the kth neuron in output layer is Back propagated.
Step 10: Finally give the selected feature and classification results.

V. RESULTS AND DISSCUSSION

Experiments are done by a personal computer with a configuration: Intel (R) Core (TM) i3-2120 CPU @ 3.30GHz, 4GB memory, Windows 7, MySQL 5.1 backend database and Jdk 1.8. The application is desktop application used tool for design code in Eclipse. Some functions used in the algorithm are provided by list of jars like weka,opencv jars etc.



	Naïve Bayes	Convolution Neural Network
Precision	69.05	79.19
Recall	81.44	63.64
F-Measure	75.11	79.31
Accuracy	81.02	89.11

VI. CONCLUSIONS

Sanskrit character identification using a Neural Network is reported in this research. Sanskrit names can be recognised, documents can be read and Sanskrit documents can be converted into structured text. 98 percent of the time, the results are correct. Character recognition and translation into other languages are possible extensions of this project. This translation from Sanskrit to English was done in English.

REFERENCES

[1] Hinton, Geoffrey E., Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R. Salakhutdinov. "Improving neural networks by preventing coadaptation of feature detectors." arXiv preprint arXiv:1207.0580 (2012).

DOI: 10.48175/568

[2] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." In Advances in neural information processing systems, pp. 1097-1105. 2012.

[3] LeCun, Yann, Bernhard Boser, John S. Denker, Donnie Henderson, Richard E. Howard, Wayne Hubbard, and Lawrence D. Jackel. "Backpropagation applied to handwritten zip code recognition." Neural computation 1, no. 4 (1989): 541-551.

[4] Nair, Vinod, and Geoffrey E. Hinton. "Rectified linear units improve restricted Boltzmann machines." In Proceedings of the 27th international conference on machine learning (ICML-10), pp. 807-814. 2010..

[5] Szegedy, Christian, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. "Going deeper with convolutions." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1-9. 2015.