Malayalam Handwritten Character Recognition

Priya P[1], Aneesh M. Haneef [2]

[1]M.tech Cse Department MES Collage of Engineering, Kuttipuram Kerala, India
[2]Assistant Professor, Department of Computer Science Engineering, Kuttippuram, Kerala, India

---***---

Abstract - —*Convolutional neural network(CNN)-based deep learning architectures are the state-of-the-art in image-based pattern recognition applications. The receptive filter fields in convolutional layers are learned from training data patterns automatically during classifier learning. There are number of well-defined, well-studied and proven filters in the literature that can extract informative content from the input patterns. Utilizing scattering transform-based wavelet filters as the first-layer convolutional filters in CNN architecture. The scattering networks are generated by a series of scattering transform operations. The scattering coefficients generated in first few layers are effective in capturing the dominant energy contained in the input data patterns. The present work aims at replacing the first-layer convolutional feature maps in CNN architecture with scattering feature maps. This architecture is equivalent to utilizing scattering wavelet filters as the first-layer receptive fields in CNN architecture. The proposed hybrid CNN architecture experiments the Malayalam handwritten character recognition which is one of the challenging multi-class classification problems. The initial studies confirm that the proposed hybrid CNN architecture based on scattering feature maps could per- form better than the equivalent self-learning architecture of CNN on handwritten character recognition problems*

Key Words: SVM, CNN, Preprocessing, Testing, Training

1. INTRODUCTION

Deep learning Techniques has achieved top class performance in pattern recognition tasks. These include image recognition [1, 2], human face recognition [3], human pose estimation [4] and character recognition [5, 6]. These deep learning techniques have proved to outperform traditional methods for pattern recognition. Deep learning enables automation of feature extraction task. Traditional methods involve feature engineering which is to be done manually. This task of crafting features is time consuming and not very efficient. The features ultimately determine the effectiveness of the system. Deep learning methods outshine traditional methods by automatic feature extraction. Handwritten character recognition is a difficult task as the characters usually has various appearances according to different writer, writing style and noise. Researchers have been trying to increase the accuracy rate by designing better features, using different classifiers and combination of different classifiers. These attempts however are limited when compared to CNN. CNNs can give better accuracy rates but it has some problems that needs to be addressed. Malayalam is one among the twenty two scheduled languages in India and is the official language in the state of Kerala, where more than 95 percentage people use Malayalam for communication [6]. Malayalam characters are complex due to their curved nature and there are characters which are formed by the combination of two characters. These along with the presence of 'chillu' make recognizing Malayalam characters a challenging task. Handwriting recognition is classically separated in two distinct domains: online and offline recognition. These two domains are differentiated by the nature of the input signal. For offline recognition, a static representation resulting from the digitalization of a document is available. Handwriting recognition refers to the recognition of handwritten paper documents which are optically scanned. The difficulty of recognition varies with a number of factors:

- Restrictions on the number of writers.
- Constraints on the writer: entering characters in boxes or in combs, lifting the pen between characters, observing a certain stroke order, entering strokes with a specific shape.
- Constraints on the language: limiting the number of symbols to be recognized, limiting the size of the vocabulary, limiting the syntax and/or the semantics.
- Many different applications currently exist, such as, check, form, mail or technical document processing. Whereas, online recognition systems are based on dynamic information acquired during the production of the handwriting.
- Malayalam Script: malayalam is the principal language of the South Indian State of Kerala. It belongs to the southern group of Dravidian Languages.
- Malayalam is spoken by over 50 million people. The Malayalam character set compromises of 95 characters consisting of the following character types:
- Vowels
- Consonants

- Anuswaram, Visargam and Chandrakkala
- Chillu
- Consonant signs
- Vowel signs: there are 13 vowels, 36 consonants, 5 chillu, 4 consonant signs, 12 vowel signs, numbers and rest contributing to anuswaram etc. Due to the peculiarities of the Malayalam language, developing a recognition system to recognize the variety of characters is a cumbersome process.

A variety of techniques of Pattern Recognition such as Template Matching, Neural Networks, Syntactical Analysis, Wavelet Theory, Hidden Markov Models, Bayesian Theory etc. have been explored to develop recognizers for different languages such as Latin, Chinese, Arabic etc. The proposed method uses direction feature extraction techniques and Neural Networks to distinguish characters and accomplish recognition tasks.

The main objective is to develop a handwritten Malayalam word recognition system. The two phases identified are:

i)   To recognize Handwritten Malayalam character
ii)   To develop Malayalam word recognition system. This is an attempt to use CNN to achieve better accuracy rate in Malayalam handwritten character recognition system.

2. LITERATURE SURVEY

A. State of the art

1) Script Identification of Multi-Script Documents: a Survey:
Approaching[1] Script identification is an important task in an OCR system for multi-lingual, multi-script documents. Many script identification methods have been proposed for written scripts at different levels within a document— page/paragraph level, text-line level, word level, and even character level. Compared to the field of document analysis and optical character recognition, research on script identification is still limited. A comprehensive overview on research activities in the field and focuses on the most valuable results obtained so far. The most vital processes in script identification are addressed in detail: identification and discriminating methods, features extraction (local and global) and classification. Different kinds of approaches have been developed and promising results have been achieved: the paper reports SoA performance results, methods concerning handwritten, printed and hybrid document processing, more research is necessary to meet the performance levels essential for everyday applications.

2) An Accelerometer-Based Digital Pen With a Trajectory Recognition Algorithm for Handwritten Digit and Gesture Recognition [2] :
The digital pen consists of a triaxial accelerometer, a microcontroller, and an RF wireless transmission module for sensing and collecting accelerations of handwriting and gesture trajectories. The proposed trajectory recognition algorithm consists of acceleration acquisition, signal preprocessing, feature generation, feature selection, and feature extraction [2]. With the reduced features, a PNN can be quickly trained as an effective classifier. In the experiments, here used 2-D handwriting digits and 3-D hand gestures to validate the effectiveness of the proposed device and algorithm. The overall handwritten digit recognition rate was 98 percentage, and the gesture recognition rate was also 98.75 percentage. This result encourages to further investigate the possibility of using our digital pen as an effective tool for HCI applications.

3) Air-Writing Recognition—Part I: Modeling and Recognition of Characters, Words, and Connecting Motions [3] :
It is an attempt to recognize air-writing with a 6 DOF motion tracking system. The writing motion is tracked with the position and orientation in the global frame, and the acceleration and angular speed in the device-wise coordinates. The air-writing recording process is very time consuming. To make the recording process feasible, place constraints on stroke orders and uppercase letters with limited vocabulary to refine the scope of air-writing data acquisition without losing too much generality. From these motion data, here derive five basic features for observations of HMMs and form the combination of pure optical, pure inertial, and complete 6-DOF features. A user study investigates input speed, motion footprint, physical strain, and subjective evaluation of two motion-based text input methods: airwriting and virtual keyboard. The results suggest that airwriting is suitable for short and infrequent text input on a motion-based user interface.

4) Telugu Handwritten Character Recognition using Zoning Features [4]:
Character recognition is one of the oldest applications of pattern recognition. Recognizing Hand-Written Characters (HWC) is an effortless task for humans, but for a computer it is a difficult job. Research in character recognition is very popular for various potential applications such as in banks, post offices, defense organizations, reading aid for the blind, library automation,

language processing and multi-media design. Optical Character Recognition (OCR) is based on optical mechanism which consists of a machine to recognize scanned and digitized character automatically. Automatic recognition of handwritten text can be done either Offline or Online. Offline handwritten recognition is the task of recognizing the image of a hand written text, in contrast to Online recognition where the dynamic characteristics of the writing are available and recorded while the scriber is writing on a special screen with a pen/stylus made for this application. Zonal based feature extraction is used in the present proposed method. The character image is divided into predefined number of zones and a statistical feature is computed from each of these zones. Usually, this feature is based on the pixels contained in that zone

3. PROPOSED METHOD

There is no standard dataset for handwritten Malayalam characters. CNN requires a large set of training images. CNN achieves a high accuracy rate only if it is trained with a substantially large training set. This is one of the biggest challenges given the time period is short. However from literature survey, it is clear that there are techniques that can be applied to increase the number of dataset images. Overall architecture of the proposed system is shown in figure 3.1. The input is first scanned using a scanner or taken as a photograph using a smart phone. The kernel weights are initialized using Gaussian distribution
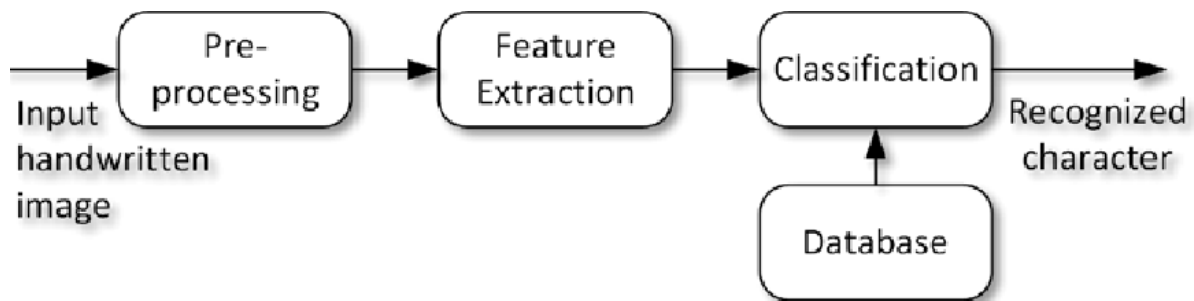

Fig -1: A Typical Handwritten Recognition System

The word is divided into different segments. The characters are written in separate panels. The features are extracted and given as input to a neural network. The characters are identified. The identified characters are obtained and are checked for word. A database of different words is stored. The written word is checked in the database and the appropriate unicode of the characters are retrieved.
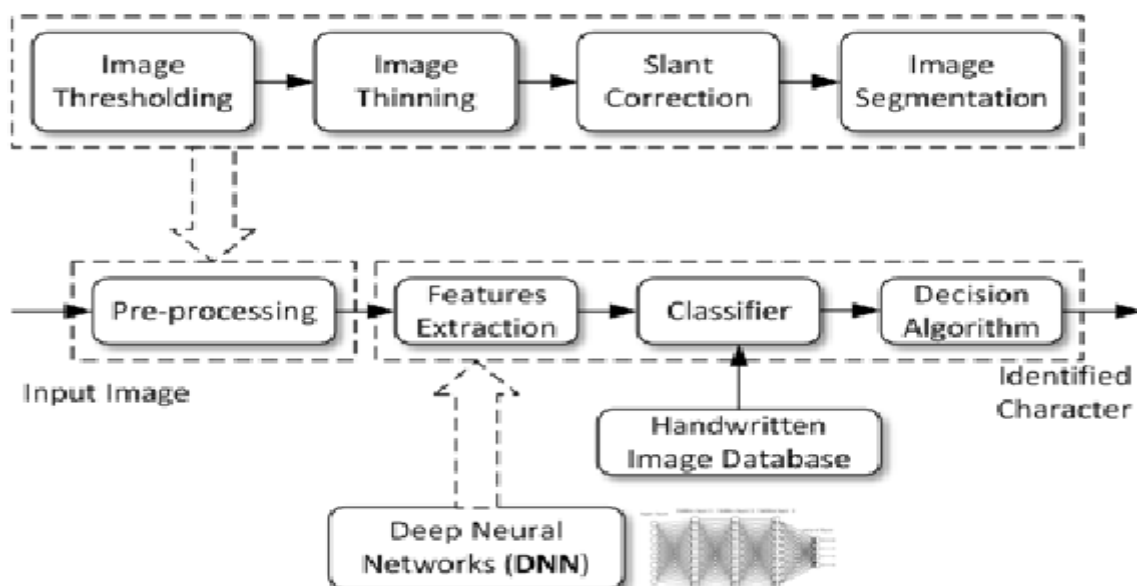
A. *System Architecture*

Fig -2: A Handwritten Character Recognition System

The entire system is divides into different modules. The various modules identified in character recognition are:

1) *Data Preprocessing :* It is a technique that is used to convert the raw data into a clean data set. The preprocessing provide the acquired data to a suitable form for further processing. In this phase the input image is generally cleaned from noise and error caused by the acquisition process. A great number of well-defined algorithms for signal processing are currently used during the preprocessing phase. However, in handwriting recognition, the preprocessing deals with more specific problems than in other fields of pattern recognition. For example, the binarization (thresholding) of the image. In preprocessing noise detection and normalization is done.

B. Feature extraction

It is the process of transforming the raw pixel values from an image, to a more meaningful and useful information. Feature extraction is defined as the problem of extracting from the raw data the information, which is most relevant for classification purpose. In this sense of minimizing within the class pattern variably while enhancing the class pattern variability. It should be clear that different feature extraction methods fulfill these requirements to a varying degree, depending on the specific recognition problem and the available data. A feature extraction method that proves to be successful in one application domain may turn out to be not very useful in another domain. Selection of feature extraction methods is probably a single most important factor in achieving high recognition performance. In addition the performance also depends on the type of classifier used. Different feature types may need different type classifiers. Also the choice of feature extraction methods limits or dictates the nature and output of preprocessing steps. Some feature extraction method work on grey level sub images of single characters, while other work on solid four or eight connected symbols segmented from the binary raster image, thinned symbols, skeletons or symbol contours.

C. Classification

A classifier is a hypothesis or discrete-valued function. It is used to assign (categorical) class labels to particular data points.The final layer of the CNN is a Softmax layer and this softmax layer is used for classifying the given input image.

D. Deep neural network

A deep neural network is a neural network with a certain level of complexity, a neural network with more than two layers. Deep neural networks use sophisticated mathematical modeling to process data in complex ways. A neural network, in general, is a technology built to simulate the activity of the human brain – specifically, pattern recognition and the passage of input through various layers of simulated neural connections. Many experts define deep neural networks as networks that have an input layer, an output layer and at least one hidden layer in between. Each layer performs specific types of sorting and ordering in a process that some refer to as "feature hierarchy." One of the key uses of these sophisticated neural networks is dealing with unlabeled or unstructured data. The phrase "deep learning" is also used to describe these deep neural networks, as deep learning represents a specific form of machine learning where technologies using aspects of artificial intelligence seek to classify and order information in ways that go beyond simple input/output protocols.

E. Decision algorithm

It is an algorithmic approach that identifies ways to split a data set based on different conditions. Decision Tree Analysis is a general, predictive modelling tool that has applications spanning a number of different areas. In general, decision trees are constructed via an algorithmic approach that identifies ways to split a data set based on different conditions. It is one of the most widely used and practical methods for supervised learning. Decision Trees are a non-parametric supervised learning method used for both classification and regression tasks. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. A decision tree is a tree-like graph with nodes representing the place where we pick an attribute and ask a question, edges represent the answers the to the question; and the leaves represent the actual output or class label. They are used in non-linear decision making with simple linear decision surface. Decision trees classify the examples by sorting them down the tree from the root to some leaf node, with the leaf node providing the classification to the example. Each node in the tree acts as a test case for some attribute, and each edge descending from that node corresponds to one of the possible answers to the test case. This process is recursive in nature and is repeated for every subtree rooted at the new nodes

F. Image thresholding

It is a simple, yet effective, way of partitioning an image into a foreground and background. This image analysis technique is a type of image segmentation. That isolates objects by converting gray scale images into binary images. Image thresholding is most effective in images with high levels of contrast. Common image thresholding algorithms include histogram and multi-level thresholding. Image analysis involves processing an image into fundamental components to extract meaningful

information. Image analysis can include tasks such as finding shapes, detecting edges, removing noise, counting objects, and calculating statistics for texture analysis or image quality. Image analysis is a broad term that covers a range of techniques that generally fit into these subcategories:

– Image enhancement to prepare images for display or analysis.
– Image segmentation to isolate regions and objects of interest.
– Noise removal using morphological filtering or deep learning.
– Region analysis to extract statistical data.

G. Image Thinning

It is a morphological operation that is used to remove selected foreground pixels from binary images. Thinning is an important step in image processing where we minimize (reduces) the width of the image i.e. the no. of pixels that helps in reducing the memory usage. The newly proposed algorithm preserves the connectivity of the image pattern and provides a reduced image after processing the given image. The algorithm is composed of three passes to get thinned image that is connected, reduced and gives clear view of whole image. Thinning is a way to reduce the size of an image which helps in minimization of memory usage of the images during their processing by the machine. It is a mechanism of removing extra pixels in order to make it (image) just one pixel thick that reduces its size. The main purpose of thinning is to reduce the data amount required for processing an image to reduce the time taken for processing of the image and easy analysis of the patterns. It is an important image analysis step for its easy analysis. After thinning the image obtained is the binary image that we take into consideration but just a pixel thick.

H. Slant correction

It is a shear transformation that attempts to make all the vertical strokes erect. Slant handwriting method to detect geometry distorted images of handwriting is being proposed.

I. Image segmentation

In computer vision, image segmentation is the process of partitioning a digital image into multiple segments (sets of pixels, also known as image objects). The goal of segmentation is to simplify and/or change the representation of an image into something that is more meaningful and easier to analyze. Image segmentation is typically used to locate objects and boundaries (lines, curves, etc.) in images. More precisely, image segmentation is the process of assigning a label to every pixel in an image such that pixels with the same label share certain characteristics. The result of image segmentation is a set of segments that collectively cover the entire image, or a set of contours extracted from the image (see edge detection). Each of the pixels in a region are similar with respect to some characteristic or computed property, such as color, intensity, or texture. Adjacent regions are significantly different with respect to the same characteristics. When applied to a stack of images, typical in medical imaging, the resulting contours after image segmentation can be used to create 3D reconstructions with the help of interpolation algorithms like Marching cubes.

4. EXPERIMENTS AND RESULTS

Malayalam handwritten character recognition is challenging, due to the strong structural resemblance among character classes and due to the presence of a large number of character classes. The proposed hybrid CNN architecture (ScatCNN) is evaluated in the context of Malayalam handwritten character recognition. ScatCNN could achieve better performance compared to the equivalent CNN architecture in validation loss and recognition accuracy.

```
*Python 3.7.5 Shell*
File  Edit  Shell  Debug  Options  Window  Help
Python 3.7.5 (tags/v3.7.5:5c02a39a0b, Oct 15 2019, 00:11:34) [MSC v.1916 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license()" for more information.
>>>
================= RESTART: C:\Users\user\Downloads\cnn (2).py =================
1
2
3
4
5
6
7
8
9
10
[76, 281, 394, 115, 113, 311, 328, 394, 146, 181]

================= RESTART: C:\Users\user\Downloads\cnn (2).py =================
1
2
3
4
5
6
7
8
9
10
[378, 345, 261, 195, 91, 11, 136, 70, 224, 65]

================= RESTART: C:\Users\user\Downloads\cnn (2).py =================
1
2
3
4
5
6
7
8
9
10
[289, 79, 108, 401, 114, 146, 33, 233, 123, 54]
```

Fig -3: Reading the dataset and preparing it for basic processing
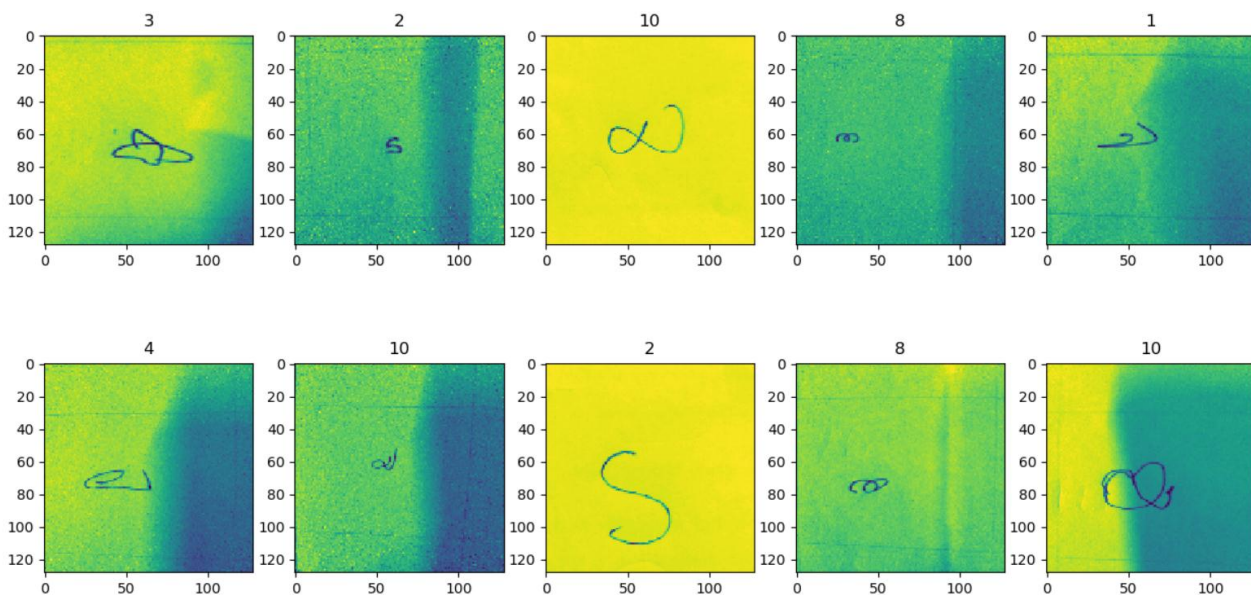

Fig -4: Result after Testing Dataset

5. CONCLUSION

Malayalam handwritten character recognition which is one of the challenging multi-class classification problems. Robust invariant feature descriptors or multistage classification techniques can solve this problem to a certain extent.

REFERENCES

[1] K. Ubul, G. Tursun, A. Aysa, D. Impedovo, G. Pirlo, T. Yibulayin, " "Script Identification of Multi-Script Documents: a Survey",, IEEE, 2017. .

[2] Jeen-Shing Wang, Fang-Chen Chuang, "An Accelerometer-Based Digital Pen With a Trajectory Recognition Algorithm for Handwritten Digit and Gesture Recognition", IEEE Transactions on industrial electronics, VOL. 59, NO. 7, JULY 2012.

[3] Mingyu Chen, Ghassan AlRegib, Biing-Hwang Juang, "AirWriting Recognition—Part I: Modeling and Recognition of Characters, Words, and Connecting Motions", IEEE Transactions on human-machine systems, 2016.

[4] Panyam Narahari Sastry, T.R.Vijaya Lakshmi, N.V.Koteswara Rao, T.V.Rajinikanth, Abdul Wahab, "Telugu Handwritten Character Recognition using Zoning Features", IEEE, 2014.

[5] K. Manjusha, M. Anand Kumar, K. P. Soman, "Integrating scattering feature maps with convolutional neuralnetworks for Malayalam handwritten character recognition", International Journal on Document Analysis and Recognition (IJDAR),2018.

[6] Manju Manuel and Saidas S. R, "Handwritten Malayalam Character Recognition using Curvelet Transform and ANN", International Journal of Computer Applications (0975 8887) Volume 121 No.6, July 2015.