INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & MANAGEMENT

MANUSCRIPT ANALYSIS AND EVENT EXTRACTION USING NATURAL LANGUAGE PROCESSING AND CONVOLUTIONAL NEURAL NETWORKS

Arun Vignesh .M[*1], Aarthe Jayaprakash[2] & Krishnap Priya .G .B[3]
[*1,2&3]Dept. Of Computer Science and Engineering, SSN college of Engineering, Chennai

ABSTRACT

The proposed system explained in this paper, deploys Natural Language Processing which intents to process and understand native and natural languages in penned form. Here we designed a system to perceive characters and words in unformatted style to a formatted pattern similar to that of a Character Recognition structure. With the aim of finding the combination of feature extraction methods for character recognition of Tamil dialects and scripts, we present, in this paper, our experimental study on feature extraction methods for character recognition on pre-civilization manuscripts. We investigated and evaluated the performance of 10 feature extraction methods and we proposed the proper and robust combination of feature extraction methods to increase the recognition rate. This system will be conducive to read age-old transcripts, manuscripts etc. and shape them into present-day dialect for people to have an easy interpretation and analysis of the pre-civilization period.

Keywords: *Natural Language Processing, Deep Learning, TensorFlow, Neural Networks, Character recognition, Convolutional Neural Networks.*

I. INTRODUCTION

During the last few years, deep learning techniques have become one of the most popular methods for character recognition. Natural Language Processing is the process of interactions between Natural Languages and the machines. This area of Artificial Intelligence concerned with processing the human beings language Tamil into a formatted form using the TensorFlow tool which is a neural network for machine translation. Thus the computer system requires to iteratively perform calculations to determine patterns by itself and translate it into digitalized form. The text can also be translated into different languages for the users' convenience. However, only a few system are available in the literature for other Asian scripts recognition. For example, some of the works are for Devanagari script [6,14], Gurmukhi script [5,15–17], Bangla script [10], and Malayalam script [18]. Those documents with different scripts and languages surely provide some real challenges, not only because of the different shapes of characters, but also because the writing style for each script differs: shape of the characters, character positions, separation or connection between the characters in a text line.

Choosing efficient and robust feature extraction methods plays a very important role to achieve high recognition performance in an IHCR and OCR [5]. The performance of the system depends on a proper feature extraction and a correct classifier selection [10]. With the aim of developing an IHCR system for Tamil script on manuscript images, we present, in this paper, our experimental study on feature extraction methods for character recognition of Tamil script. It is experimentally reported

that to improve the performance of an IHCR system, the combination of multi features is recommended. Tamil is one of the oldest languages in the world. It has a 2000 years old literary works which have undergone series of transformations in its formats. It is considered to be the National Language of many countries and official state language of states and union territories. In this system, we recognize the characters from these old writings and translate them into the present-day understandable format using different algorithms and tools in an efficient and novel way.

II. LITERATURE SURVEY

The paper authored by A. Shaina Gupta et.al. [1] (2014) focuses mainly on Natural Language Processing and finding the words in a document and translating them from one language to another via Optical Character Recognition. Recognizing handwritten document is an ambitious task. The problems which are encountered while

character recognition is broken character, heavy printing, spacing problem, shape variance etc. Therefore the process of deriving foreground and background from the image document is done. Thus the Binarization algorithm is deployed. The character image is created by reading the image, preprocessing it into a logical format, binarization, resizing the image and applying zoning algorithm. Thus the feature extraction and characterization are carried out in detail.

The work carried out by B.Vishnu Sundaresan et.al [2] (2015) gears the problem of understanding natural handwritten images. It includes learning algorithms like K nearest neighbors, support vector machine, and stochastic gradient to segregate the images. The slope and slant estimation is done for simpler segmentation. The average thickness of the slope is also calculated. A Gabor filter for edge detection is used. The learning rate, dropout rate, MNIST accuracy, segmentation area done. By slant correction, a better picture of the characters can be obtained. The data augmentation is done and characters are recognized.

This system provides a novel handwritten English recognition system was proposed by C. Aiquan et.al [3] (2012) . The original sample undergoes preprocessing and segmentation of the characters are done. A character hypothesis is generated which is then recognized to discover the result. The results are scored to find the best results. The fragmentary segments in multi-strokes and intimate segments are merged. The character recognition is followed by the word recognition. The scoring and sorting of these results are necessary. Dynamic Pruning is deployed to find the word which is a traversal of combine-tree. Thus this paper explores segmentation techniques by calculating the hamming distance and EC codes.

The paper authored by D.Pranav et.al [4] (2017) explains the conversion of the input text into formatted code using OCR is the basic principle. The handcrafted features are used to find the character sets. The undesirable entities from the characters are removed and the data set is created for the language. Dataset augmentation is done to train the e CNN. Affine transformation (Translation, Scaling, Rotation, and Sheering) is used for augmentation. The CNN networks are tuned and the input text images are given which is then classified and the output is obtained. Hence the system gives higher accuracy rate.

III. TAMIL CHARACTER SET RECOGNITION

A. Character set

The Tamil character set consists of 18 consonants, 12 vowels, and one special character. The complete script, therefore, consists of the 31 letters in their independent form and an additional 216 combinations of letters, for a total of 247 combinations of a consonant and a vowel, a mute consonant, or a vowel alone. These combinations of letters are formed by adding a vowel marker to the consonant. Some vowels require the basic shape of the consonant to be altered in a way that is specific to that vowel. While others are written by including a vowel-specific suffix to the consonant, yet allow others a prefix, and still, other vowels require including both a prefix and a suffix to the consonant. In every case, the vowel marker is different from the standalone character for the vowel

ISSN 2277 – 5528
Impact Factor- 4.015

Fig 1: Old Tamil Characters

B. Handwritten Documents

A manuscript is a book or document written using hand rather than printed. Recognizing handwritten documents is a difficult task as there will be a difference in handwriting and also breakage of numerals, letters, shaping problems etc

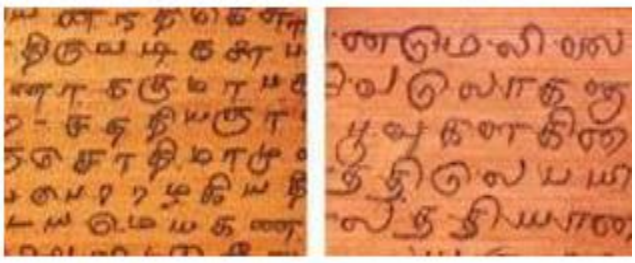

Fig 2: Handwritten Tamil Manuscripts

C. Numerals

The Tamil numerals has its symbolic form.

Fig 3: Tamil Numerals Representation

ISSN 2277 – 5528
Impact Factor- 4.015

IV. TECHNICAL APPROACH

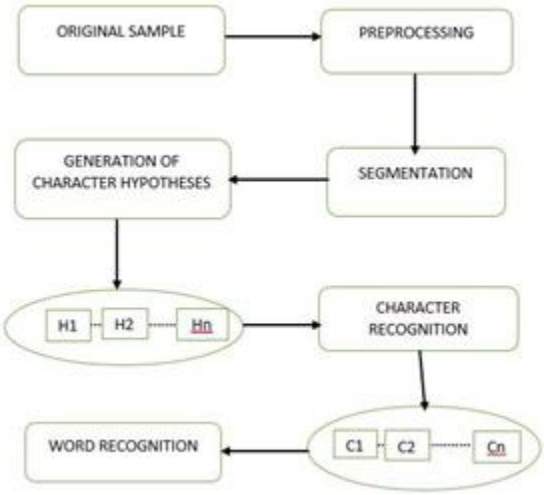

Fig 4: Flow of Operations

A. Character Recognition

For training the MNIST datasets we make use of many convolutional neural network architectures and compare its accuracy with other architectures. The characters are identified using electric devices or computer software called as Optical Character Recognition. On moving to the character datasets, we plan on implementing them using supervised classifier models and attain accuracy as well as incorporate deep convolutional network algorithms. The characters must be trained one at a time. Advanced systems which can produce a high degree of recognition accuracy of fonts are common with support for different types of digital image file format inputs. Certain software prototypes are capable of reproducing formatted output that is similar to the original sample including images, tables, and also other non-textual components. The dark or light areas in the scanned-in image or bitmap are analyzed in order to identify each alphabet or number. Once a character is recognized, it is converted into an ASCII code. Special circuit boards and computer chips are used to fasten up the recognition process.

B. Segmentation

Slope and slant correction values are calculated for a simpler process. The thickness of the words is found and smoothed. The size of each character is found and the standard slant angle is set.

Fig 5: Slant Correction

1. Preprocessing

The characters are preprocessed to remove the undesirable entities or any leakage of ink etc. for clearer view of the characters. Slope and slant corrections are much needed to reduce the unnecessary variations in handwritten samples. The image is resized into suitable form and padding is done i.e. white spaces are added for clear view of the characters. Careful calculation of slope and slant can make the segmentation process to be carried out easily and in an efficient way. In preprocessing a core region is identified by measuring between the horizontal line and the slanted characters. It also includes an estimation of the average thickness of the stroke, which can be used in slant detection and correction.

2. Word recognition

After character recognition, we have the recognition results for character hypotheses of one word. Extracting the recognition results for the word from the recognition results of character hypotheses is the next step. Choose the vertical projection of each image to identify the segmentation points to determine and realize words. The vertical projection is determined using the addition of all white pixels along a line in the vertical direction, which is accomplished by passing the aggregate thickness of each stroke to the algorithm and using them as a cutoff threshold. Meanwhile, with more than one recognition result, it is quite essential to calculate the possibilities of them to be the label of the word. Moreover, scoring and sorting of these recognition results are also necessary for TOP X results.

3. Feature Extraction Methods and Proposed combination of features

Many feature extractions methods have been presented in the literature. Each method has its own advantages or disadvantages over other methods. In addition, each method may be specifically designed for some specific problem. Most of feature extraction methods, extract the information from binary image or gray scale image . Some surveys and reviews on features extraction methods for character recognition were already reported. In this work, first, we investigated and evaluated some most commonly used features for character recognition: histogram projection, celled projection, distance profile, crossing, zoning, moments, some directional gradient based features, Kirsch Directional Edges, and Neighbourhood Pixels Weights. Secondly, based on our preliminary experiment results, we proposed and evaluated the combination of NPW features applied on Kirsch Directional Edges images, with Histogram of Gradient (HoG) features and zoning method. This section will only briefly describe the feature extraction methods which were used in our proposed combination of features and the convolutional neural network. For more detail description of other commonly used feature extraction methods which were also evaluated in this experimental study, please refer to references mentioned above.

A. Our proposed combination of features

After evaluating the performance of 10 individual feature extraction methods, we found that the HoG features, NPW features, Kirsch features and Zoning method give a very promising result. We obtained 62,45% of recognition rate only by using Kirsch features. It means that the four directional Kirsch edge images already serve as a good feature discriminants for our dataset. The shape of Balinese characters are naturally composed by some curves. We can notice that Kirsch edge image is able to give the initial directional curve features for each character. On the other hand, NPW features have an advantage that it can be applied directly to gray level images. Our hypothesis is the four directional Kirsch edge images will provide a better feature discriminants for NPW features. Based on this hypothesis, we proposed a new feature extraction method by applying NPW on kirsch edge images. We call this new method as NPW-Kirsch . Finally, we concatenate NPW-Kirsch with two other features, HoG and Zoning method.

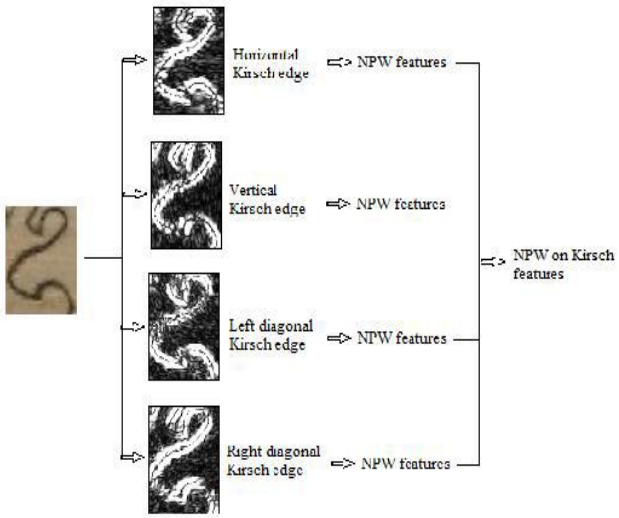

Fig.6 Scheme of NPW on Kirsch Features

B. Convolutional neural network

In this experiment, we also use convolutional neural network where feature extraction step is not required. The network considers the value of each pixel of the input image as the input layer. In this experiment, we used Tensorflow library2.

More specifically, multilayer convolutional neural network is used. The architecture of our network is illustrated in Fig. 8. Given an input gray scale image of 28x28 pixels, the first convolutional layer (C1) is computed by using a sliding window of 5x5 pixels. We obtain C1 layer containing 28x28x32 neurons, where 32 is the number of features maps chosen. For each sliding window on the neuron (i, j), the convolutional C(i, j) output can be computed by:

$$C(i,j) = \sum_{k=0}^{4} \sum_{l=0}^{4} W_{(k,l)} I_{(i+k,j+l)}$$

where W is a 5x5 matrix to be used as the shared weights matrix, and I is the input neurons. Rectified linear unit is then applied. We obtain: C=ReLu(b+C), where b is the bias. Then, we apply max-pooling using a window of 2x2 which consists in choosing the maximum activation in the selected region as the output. We obtain P2 layer consists of 14x14x32 neurons. After computing the second convolutional layer (C3) and second max-pooling (P4), we obtain a layer (P4) of 7x7x64 neurons. We add a fully-connected layer (F5) of 1024
neurons, where

$F5 = \text{Re}\,Lu(P4W4 + b4)$

P'4 is a one dimension matrix containing the 3136 neurons in P4. W4 is a 3136x1024 shared weight matrix, and b4 is the bias.
Finally, we fully connect this F5 layer to the output layer, where the number of neurons equals the number of classes by using the equation

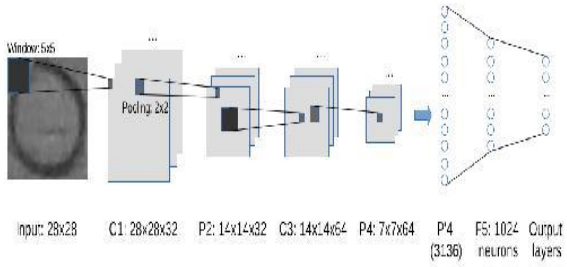

Fig.7 Architecture of multilayer convolutional neural network

C. Experiments of character recognition

We investigated and evaluated the performance of 10 feature extraction methods and the proposed combination of features in 29 different schemes. We also compared the experimental result with the convolutional neural network. For all experiments, a set of image patches containing tamil characters from the original manuscripts will be used as input, and a correct class of each character should be identified as a result. We used k=5 for k-NN classifier, and all images are resized to 50x50 pixels (the approximate average size of character in collection), except for Gradient features where images are resized to 81x81 pixels to get evenly 81 blocks of 9x9 pixels, as described in. The results (Table I) show that the recognition rate of NPW features can be significantly increased (up to 10%) by applying it on the four directional Kirsch edge images (NPWKirsch method). Then, by combining this NPW-Kirsch features, HoG features and Zoning method can increase the recognition rate up to 85%. This result is slightly better than using the convolutional neural network. In our experiments, the number of training dataset for each classes is not balance, and it influences the performance of convolutional neural network. But this condition was already clearly stated and can not be avoided as real challenge of our IHCR development for pre-civilization Tamil scripts on manuscripts. Some ancient characters are not frequently found in our collection of manuscripts.

ISSN 2277 – 5528
Impact Factor- 4.015

Table 1. Recognition rate from all schemes of experiment

S.No.	Method	Feature Dim.	Classifier	Recog Rate %
1	Histogram Projection (Binary)	100	SVM	26.313
2	Celled Projection (Binary)	500	SVM	49.941
3	Celled Projection (Binary)	500	K-NN	76.161
4	Distance Profile (Binary)	200	SVM	40.127
5	Distance Profile (Binary)	200	K-NN	58.947
6	Distance Profile (Skeleton)	200	SVM	36.765
7	Crossing (Binary)	100	SVM	15.007
8	Zoning (Binary)	205	SVM	50.645
9	Zoning (Binary)	205	K-NN	78.535
10	Zoning (Skeleton)	205	SVM	41.848
11	Zoning (Grayscale)	205	SVM	52.417
12	Zoning (Grayscale)	205	K-NN	66.128
13	Gradient Feature (Gray)	400	SVM	60.041
14	Gradient Feature (Gray)	400	K-NN	72.579
15	Moment Hu (Gray)	56	SVM	33.481
16	Moment Hu (Gray)	56	K-NN	33.481
17	HoG (Gray)	1984	SVM	71.275
18	HoG (Gray)	1984	K-NN	84.347
19	NPW (Binary)	100	SVM	51.388
20	NPW (Gray)	100	SVM	54.129
21	Kirsch (Gray)	100	SVM	62.452
22	HoG with Zoning (Gray)	1984	SVM	69.685

23	HoG with Zoning (Gray)	1984	K-NN	83.501
24	NPW-Kirsch (Gray)	400	SVM	63.573
25	NPW-Kirsch (Gray)	400	K-NN	76.711
26	HoG on Kirsch edge (Gray)	1984*4	K-NN	82.093
27	HoG + NPW-Kirsch (Gray)	1984+400	K-NN	84.752
28	Zoning + Celled Projection (Binary)	205+500	K-NN	77.701
29	HoG + NPW-Kirsch (Gray) + Zoning (Binary)	1984+400+205	K-NN	85.156
30	Convolutional Neural Network			84.308

V. TENSORFLOW

The TensorFlow tool is used for deep learning technique. It allows a user to train the computer system to perform a specific task by feeding a large amount of data. The characters and numbers are fed to train the system to recognize it. A graph of data flows can be expressed using TensorFlow. Data in TensorFlows are represented as multidimensional arrays. Thus using this neural network will speed up the process and also affordable by modern hardware.

VI. AUGMENTED REALITY

The augmented reality platform combines the line between what is real and what is done by the machine. It elevates what we look, hear and smell. A computer generated image overlays on the view of the real world. Here we can use either a Goggles to witness the recognized characters through augmented reality platform or build an AR application for the same.

VII. CONCLUSION

Thus this system is a novel method for character recognition of artifacts. The TensorFlow tool makes it easier for training the characters using deep learning classifier algorithms. The time consumed are less when compared to its counterparts. The unwanted qualities of the image is cleared for better understanding. We proposed the proper and

robust combination of feature extraction methods to increase the recognition rate. Our study shows that the recognition rate can be significantly increased by applying NPW features on four directional Kirsch edge images and the use of NPW on Kirsch features in combination with HoG features and Zoning method can increase the recognition rate up to 85%, and it still slightly better than using the convolutional neural network. A mobile application is created for viewing it in an Augmented Reality platform. It is advantageous in pattern recognition close to artificial intelligence.

REFERENCES

[1] Shaina Gupta , International Journal of Computer Science and Mobile Computing "Recognition of Handwritten Devnagari numerals with SVM Classifier" .
[2] Vishnu Sundaresan, Jasper Lin, "Recognizing Handwritten Digits and Characters".
[3] Aiquan Yuan, Gang Bai, Po Yang, Yanni Guo, Xinting Zhao, 2012 International Conference on Frontiers in Handwriting Recognition.Handwritten EnglishWord Recognition based on Convolutional Neural Network.
[4] Pranav P Nair, Ajay James, C Saravanan, International Conference on Inventive Communication and Computational Technologies, "Malayalam Handwritten Character Recognition Using Convolutional Neural Network."
[5] Study on Feature Extraction Methods for Character Recognition, **2016 23rd** International Conference on Pattern Recognition (ICPR) Cancún Center, Cancún, México, December 4-8, 2016.
[6] S. Kumar, Neighborhood Pixels Weights-A New Feature Extractor, Int. J. Comput. Theory Eng. (2009) 69–77. doi:10.7763/IJCTE.2010.V2.119.
[7] N. Arica, F.T. Yarman-Vural, Optical character recognition for cursive handwriting, IEEE Trans. Pattern Anal. Mach. Intell. 24 (2002) 801–813. doi:10.1109/TPAMI.2002.1008386.
[8] M. Blumenstein, B. Verma, H. Basli, A novel feature extraction technique for the recognition of segmented handwritten characters, in: IEEE Comput. Soc, 2003: pp. 137–141. Doi : 10.1109 / ICDAR. 2003. 1227647.
[9] M. Bokser, Omnidocument technologies, Proc. IEEE. 80 (1992) 1066–1078. doi:10.1109/5.156470.
[10] M. Zahid Hossain, M. Ashraful Amin, Hong Yan, Rapid Feature Extraction for Optical Character Recognition, CoRR. abs/1206.0238 (2012). http://arxiv.org/abs/1206.0238.
[11] Y. Fujisawa, Meng Shi, T. Wakabayashi, F. Kimura, Handwritten numeral recognition using gradient and curvature of gray scale image, in: IEEE, 1999: pp. 277–280. doi:10.1109/ICDAR.1999.791778.
[12] Z. Jin, K. Qi, Y. Zhou, K. Chen, J. Chen, H. Guan, SSIFT: An Improved SIFT Descriptor for Chinese Character Recognition in Complex Images, in: IEEE, 2009: pp. 1–5. doi:10.1109/CNMT.2009.5374825.
[13] M. Rani, Y.K. Meena, An Efficient Feature Extraction Method for Handwritten Character Recognition, in: B.K. Panigrahi, P.N. Suganthan, S. Das, S.C. Satapathy (Eds.), Swarm Evol. Memetic Comput., Springer Berlin Heidelberg, Berlin, Heidelberg, 2011: pp. 302–309. http://link.springer.com/10.1007/978-3-642-27242-4_35 (accessed March 21, 2016).
[14] R.J. Ramteke, Invariant Moments Based Feature Extraction for Handwritten Devanagari Vowels Recognition, Int. J. Comput. Appl. 1 (2010) 1–5. doi:10.5120/392-585.
[15] G.S. Lehal, C. Singh, A Gurmukhi script recognition system, in: IEEE Comput. Soc, 2000: pp. 557–560. doi:10.1109/ICPR.2000.906135.
[16] D. Sharma, P. Jhajj, Recognition of Isolated Handwritten Characters in Gurmukhi Script, Int. J. Comput. Appl. 4 (2010) 9–17. doi:10.5120/850-1188.
[17] K. Singh Siddharth, R. Dhir, R. Rani, Hand written Gurmukhi Numeral Recognition using Different Feature Sets, Int. J. Comput. Appl. 28 (2011) 20–24. doi:10.5120/3361-4640.
[18] Ashlin Deepa R.N, R.Rajeswara Rao, Feature Extraction Techniques for Recognition of Malayalam Handwritten Characters: Review, Int. J. Adv. Trends Comput. Sci. Eng. IJATCSE. 3 (2014) 481–485.
[19] Ø. Due Trier, A.K. Jain, T. Taxt, Feature extraction methods for character recognition-A survey, Pattern Recognit. 29 (1996) 641–662. doi:10.1016/0031- 3203(95)00118-2.
[20] Satish Kumar, Study of Features for Hand-printed Recognition, Int. J. Comput. Electr. Autom. Control Inf. Eng. 5 (2011).
[21] Neha J. Pithadia, Dr. Vishal D. Nimavat, A Review on Feature Extraction Techniques for Optical Character Recognition, Int. J. Innov. Res. Comput. Commun. Eng. 3 (2015)..