International Journal of Computational Intelligence Research
ISSN 0973-1873 Volume 13, Number 5 (2017), pp. 933-943

http://www.ripublication.com

Handwritten Gurumukhi Character Recognition Using Convolution Neural Network

Harpreet Kaur

M. Tech. Research Scholar (Computer Science & Engineering), Yadavindra College of Engineering, Punjabi University Guru Kashi Campus, Talwandi Sabo, Bathinda, Punjab, India

Simpel Rani

Associate Professor (Computer Science & Engineering), Yadavindra College of Engineering, Punjabi University Guru Kashi Campus, Talwandi Sabo, Bathinda, Punjab, India

Abstract

Handwritten Character Recognition (HCR) is one of the challenging processes. Automatic recognition of handwritten characters is a difficult task. In this paper, we have presented a scheme for offline handwritten Gurmukhi character recognition based on CNN classifier. The system first prepares a skeleton of the character, so that feature information about the character is extracted. CNN based approach has been used to classify a character based on the three features like Zoning, Horizontal Peak Extent and Diagonal. We have taken the samples of offline handwritten Gurmukhi characters from 70 different writers. We have experimented partition strategy for selecting the training and testing patterns. We have used in all 2450 images of Gurmukhi characters for the purpose of training and testing. We have used Zoning, Diagonal and Horizontal Peak Extent feature extraction techniques in order to find the feature sets for a given character. The proposed system achieves a maximum recognition accuracy of 92.08% with 90% training data and 10% testing data using Zoning based features and CNN Classifier.

Keywords: Handwritten character recognition, Feature extraction, Zoning based Features, Diagonal features, Horizontal Peak Extent, CNN.

1. INTRODUCTION

Various published work on Indian scripts recognition deals with printed documents and very few articles deal with handwritten script problem. It has motivated us to consider the handwritten script recognition for Gurmukhi script. HCR abbreviated as Handwritten Character Recognition usually. It is the process of converting handwritten text into machine printed format. HCR can be online or offline. In online handwriting recognition, data captured during the writing process with the help of a special pen and an electronic surface. Offline documents are scanned images, generally on a sheet of paper. Offline handwriting recognition is significantly different from online handwriting recognition. We have proposed a recognition system for offline handwritten Gurmukhi characters. Recognition system consists of the activities namely, digitization, preprocessing, features extraction and classification. Last some decades HGCR is an area of pattern recognition that has been the subject of considerable research. There are number of applications (i.e. Indian offices such as bank, sales-tax, railway, embassy, etc.). We can used the both English and regional languages. Number of forms and applications are filled in regional languages and sometimes those forms have to be scanned directly. If there is no HGCR system, then image is directly captured and there is no option for editing those documents. HCR is a process of automatic computer recognition of characters in optically scanned and digitized pages of text. The main objective of this system is to recognize alphabetic Gurumukhi characters, which are in the form of digital images, without any human intervention. We have done this by searching a match between the features extracted from the given character's image and the library of image models. There are number of researchers have already worked on the recognition problem of offline printed characters. For example, a handwritten Gurumukhi character recognition by using a modified division points (MDP) feature extraction technique recognition system has been proposed by Kumar et al. [1]. Kumar et al. [2] has proposed identify or recognize the handwritten characters. Siddharth et al. [3] have proposed handwritten Gurmukhi numeral recognition using different feature sets.

2. INTRODUCTION OF GURUMUKHI SCRIPT

The word "Gurumukhi "exactly means "from the mouth of the guru". Punjabi is the world's 14th most widely spoken language. Punjabi speakers are spread over all parts of the world not only confined to north Indian states such as Punjab, Haryana etc. In this language there is rich literature in the form of scripture, books, poetry. Therefore, it is, important to develop offline handwriting recognition for such a widely used language which may find many practical uses in various areas. Gurumukhi script symbol is writing from left to right side of the paper. In Gurumukhi, there is no upper or lower case characters concept. Gurumukhi contained with 41 consonants and 12 vowels as shown in Figure 1 and Figure 2. Gurumukhi has two dimensional compositions of symbols with connected and disconnected diacritics

Figure 1: Punjabi Vowels and Vowel Diacritics (Laga Matra).

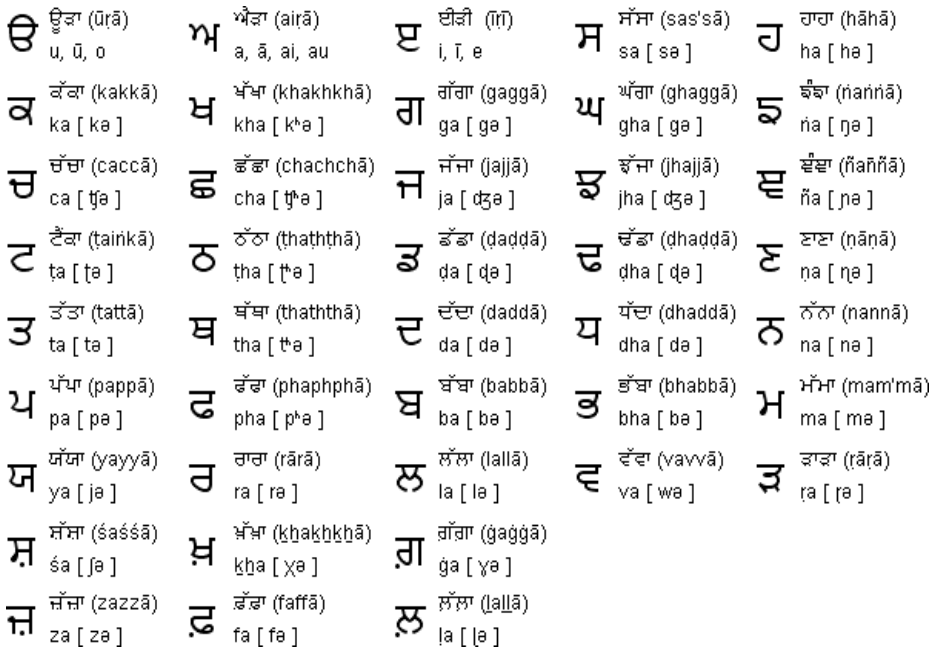

Figure 2: Punjabi Consonants.

3. DATA COLLECTION

Collection of data of Gurumukhi scripts for our implementation is collected from 70 different persons. Each writer commits to write 35 samples of different Gurumukhi characters. We take these samples on white papers written in an isolated manner. Some of the samples of our collected dataset show in Figure 3.

Script Character	W1	W2	W3	W4	W5
ੳ					
ਅ					
ੲ					
ਸ					
ਹ					

Figure 3: Collected dataset

A sample of 70 writers was selected from schools, colleges and government offices. We requested to these writers were write each Gurumukhi character. A sample of five handwritten Gurumukhi characters by five different writers (W1, W2… W5) is given in Figure 3.

4. THE RECOGNITION SYSTEM

The Recognized system consists of the different Stages like digitization, preprocessing, feature extraction and classification. The block diagram of proposed recognition system is given in Figure 4.

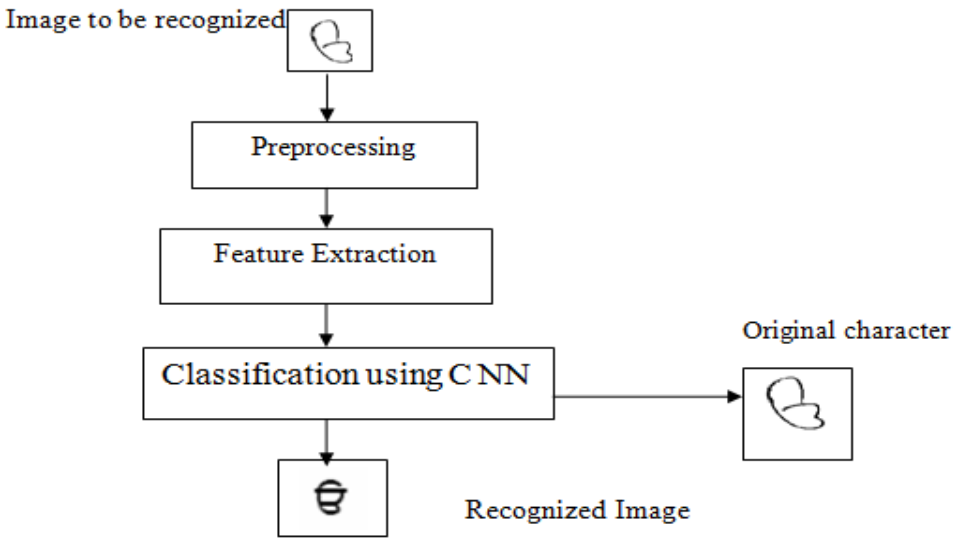

Figure 4: Diagram of Recognition of Handwritten Gurumukhi Character.

4.1 Digitization

It is the process of Converting a paper based handwritten document into an electronic form. Electronic conversion is carried out using a process wherein a document is scanned and then a bitmap image of the original document is produced.

4.2 Pre-processing

It is also the phase of handwritten character recognition system. Pre-processing includes noise removal, skew detection/correction and skeltonization. Document of Pre-processing is used to detect and remove all unwanted bit pattern which may leads to reduce the recognition accuracy.

4.3 Feature Extraction

We are using various feature extraction techniques for recognition purpose. We've got used following three sets of features extracted to understand Gurumukhi numerals. Those methods are used to apprehend Gurumukhi handwritten characters.

1. Zoning features

1. Diagonal features
2. Horizontal peak extent features

4.3.1 Zoning Features Extraction:

In zoning, the Character Image is split into N×M zones. The purpose of zoning is to attain the neighborhood traits. Zoning primarily based feature is the best method for extraction of features. In this method, digitized image is split in to n range of divisions having each of identical length. After that, we've calculation of the pixel density is performed in each area via considering the variety of foreground pixels in the corresponding region. We've taken a picture with 100×100 Pixels. Here, we've got divided the digitized image into n=100 identical region and the use of this approach we've got extracted a hundred features for each image as shown in Figure 5**.**

Z1	Z2	Z3	Z4	Z5	Z6	Z7	Z8	Z9	Z10
Z11	Z12	Z13	Z14	Z15	Z16	Z17	Z18	Z19	Z20
Z21	Z22	Z23	Z24	Z25	Z26	Z27	Z28	Z29	Z30
Z31	Z32	Z33	Z34	Z35	Z36	Z37	Z38	Z39	Z40
Z41	Z42	Z43	Z44	Z46	Z46	Z47	Z48	Z49	Z50
Z51	Z52	Z53	Z54	Z56	Z56	Z57	Z58	Z59	Z60
Z61	Z62	Z63	Z64	Z65	Z66	Z67	Z68	Z69	Z70
Z71	Z72	Z73	Z74	Z75	Z76	Z77	Z78	Z79	Z80
Z81	Z82	Z83	Z84	Z85	Z86	Z87	Z88	Z89	Z90
Z91	Z92	Z93	Z94	Z95	Z96	Z97	Z98	Z99	Z100

Figure 5: Zones of any input character

4.3.2 Diagonal features

In this method, the Character Image is divided into n divisions and features of image are extracted from the pixels by means of moving alongside its diagonal. Each division consists of (2n-1) diagonals, whose values are averaged to get a single value as a feature value of feature vector for complete image.

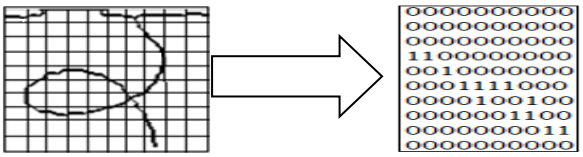

Figure6 (a): Diagonal feature extraction

Figure 6(b): Diagonals of **Z45** zone.

4.3.3 Horizontal peak extent features

We describes in this technique, sum of successive foreground pixel extents in horizontal direction in each row are considered. This technique is also gives better results. Image having size of 100×100 is divided into *n*=100 zones and therefore, 100 features have been extracted. We have been used for extract the horizontal peak extent based features in following steps:

Step 1: The input image divide into *n* numbers of divisions, each containing m×m pixels.

Step 2: Calculate the sum of successive foreground pixel extents in horizontal direction in each row of division.

Step 3: Then check the largest value in each row and swap it with each foreground pixel in row.

In Figure 6 (a) and (b), we have depicted the process of peak extent based features.

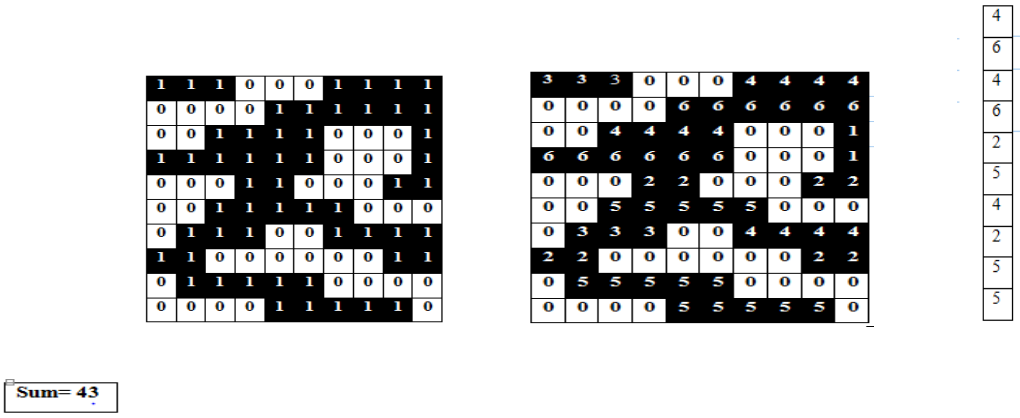

Figure 7: Peak extent based features: (a) bitmap image, (b) horizontal peak extent based feature.

4.4 Classification

It is also an important phase of handwritten character recognition system. This phase is also known as decision making phase and this phase uses the feature extracted in the previous phase, namely, feature extraction phase. The main aim of classification phase is to recognize the input data. In this work, we have considered CNN.

4.5 Convolution Neural Network (CNN)

Numerous works in handwritten character recognition are available for English with respect to Gurumukhi or other major languages. It is based on Gurumukhi handwritten character recognition is investigated. This method normalizes the written character images and then employs CNN to classify individual characters. It does not employ any feature extraction method.CNN is very much similar to ordinary Neural Networks. It made from neurons that have Learnable weights and biases. Each neuron receives some inputs, performs a dot product and optionally follows it with a non-linearity. The full network still express a single differentiable record function from the raw image pixels on one end to class scores at the other. And they still have a loss function on the last layer and all the tricks we developed for learning regular Neural Networks still apply.

5. EXPERIMENTAL RESULTS

The outcome of character recognition system for offline handwritten Gurumukhi characters are provided here. These results are based on three feature extraction techniques like zoning, diagonal and horizontal peak extent and also the combination of these three features. We have used different data set strategies. First strategy *a*, we have used 50% data for training set and 50% data in the testing set. Then strategy *b*,

we have taken 60% data for training set and 40% data used for testing set. Strategy *c* has 70% data in training set and 30% data in testing set. For strategy *d* has 80% data in training set and 20% in testing set. Strategy *e* has used 90% data in training set and remaining 10% data in testing set.

Experimental Feature-wise results are presented in the following sub-sections.

5.1 Recognition accuracy of Zoning feature extraction

In Zoning feature extraction technique, we have acquired the results by calculating the 100 features of every character. We have achieved maximum accuracy of 92.08 % by using Zoning based features. It has been achieved using CNN classifier. Graphical representation of these results has been shown in Figure 8.

Figure 8: Recognition Accuracy of Zoning Feature Extraction technique.

5.2 Recognition accuracy of Zoning and Diagonal feature extraction

We have obtained similar maximum accuracy of 87.79% by using Zoning and Diagonal based features. In this Zoning and Diagonal based features are used for the input to CNN classifier. Maximum accuracy achieved is 87.79% in 10 fold cross validation as shown in Figure 9.

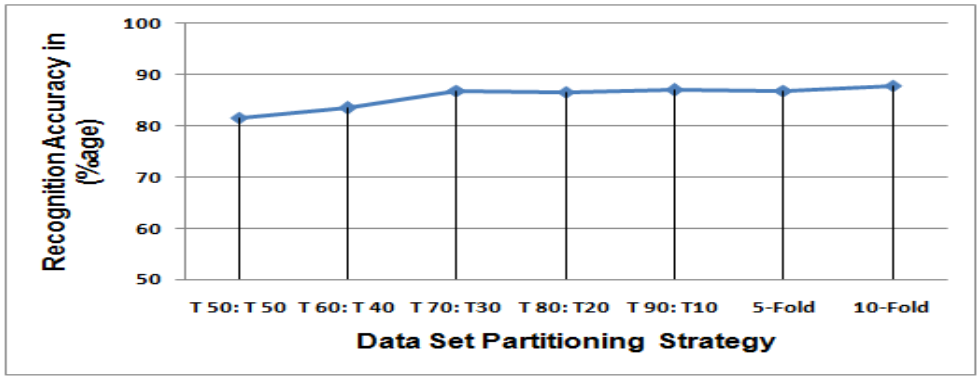

Figure 9: Recognition Accuracy Graph of zoning and diagonal feature extraction

5.3 Recognition accuracy of Diagonal feature extraction

In this method, we have achieving the highest recognition accuracy of %91.63 using 10 fold cross validation as shown in Figure 10.

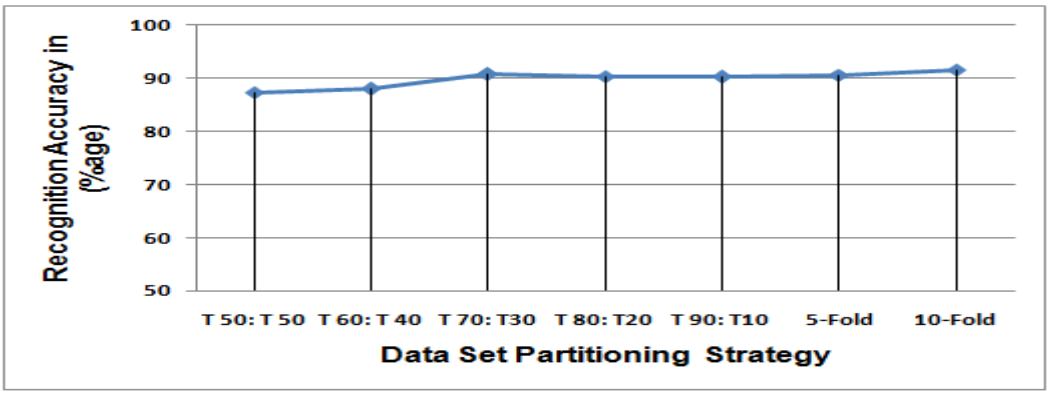

Figure 10: Recognition Accuracy Graph of Diagonal Feature extraction.

5.4 Recognition accuracy using combination of all three features

Figure 11 shows the recognition accuracy by performing the combination of zoning feature (F1), Diagonal feature (F2), peak extent feature (F4). We have obtained similar maximum accuracy of 87.22 % by using10 fold cross validation in this case. It has been achieved using CNN classifier.

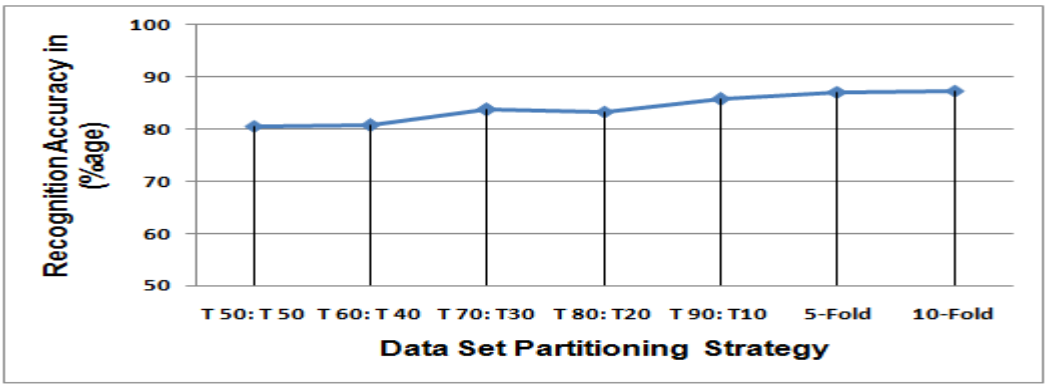

Figure 11: Recognition Accuracy using combination of all three features (F1+F2+F4).

Experiments have also been performed using F1 + F2, F1 + F4 and F2 + F4. The combined experimental results have been shown in Table 1.

Table 1: Recognition accuracy using various feature extraction techniques using CNN.

Feature Extraction Techniques	CNN
Zoning (F1)	92.08
Diagonal(F2)	91.63
Peak Extent(F4)	81.34
F1+F2	87.79
F2+F4	86.85
F1+F4	86.97
F1+F2+F4	87.34

One can see that F1 and F2 in combination are giving best accuracy of 87.79 for recognizing handwritten Gurumukhi characters using CNN. Accuracy Graph of character recognition by various Feature extraction techniques shown in Figure12.

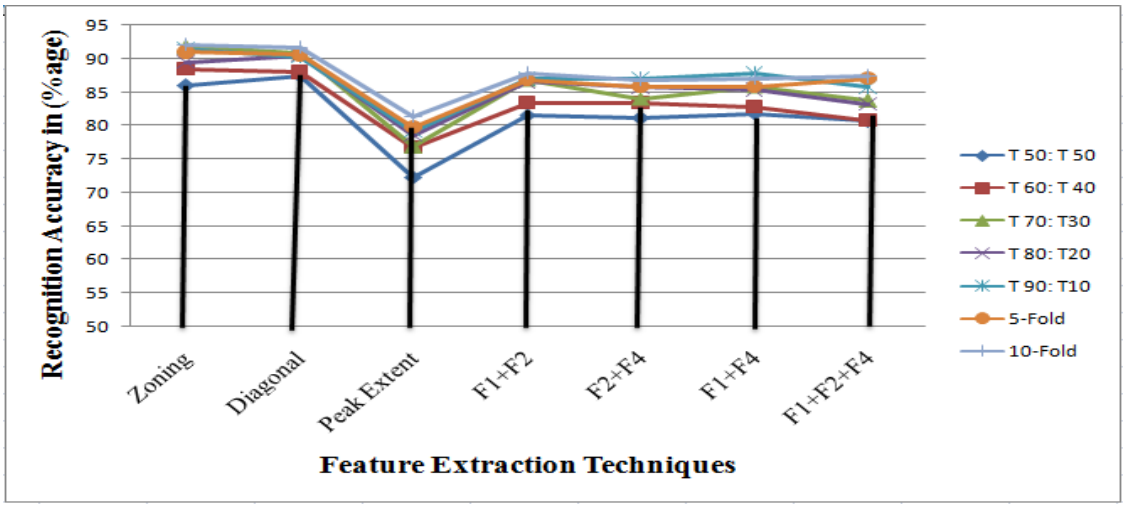

Figure 12: Character Recognition Accuracy of Various Feature Extraction Techniques.

6. CONCLUSION

Handwritten Character Recognition is one of the important step of OCR**.** We described a classifier is proposed for text recognition of complex handwritten Gurumukhi document images. Handwritten Gurumukhi script has some complexities Like Resembling characters. These cause majorities of the error during recognition stage. From the results we can say, the expected method is very useful for recognize a Gurumukhi Characters. We have achieved encouraging and satisfactory results on the complex handwritten documents. In future, we are trying to implement the good and suitable method for recognized Characters.

REFERENCES

[1] M. Kumar, M. K. Jindal and R. K. Sharma, "Classification of Characters and Grading Writers in Offline Handwritten Gurmukhi Script", Proceedings of the 2011 International Conference on Image Information Processing, IEEE, pp. 1-4, (2011).

[2] M. Kumar, R. K. Sharma and M. K. Jindal, "Offline Handwritten Gurumukhi Character Recognition: Study of Different Feature-Classifier Combinations", Workshop on Document Analysis and Recognition, pp. 94-99, (2012).

[3] D. Sharma and D. Gupta, "Isolated Handwritten Digit Recognition using Adaptive Unsupervised Incremental Learning Technique", International Journal of Computer Applications, Vol. 7, No.4, September (2010).

[4] Macwan, J. J., Goswami, M. M., & Vyas, A. N., "A survey on offline handwritten north Indian script symbol recognition", In Proceedings of International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)", pp. 2747-2752, (2016).

[5] Kumar, Munish Jindal, M. K. and Sharma, R. K., "Classification of Characters and Grading Writers in Offline Handwritten Gurumukhi Script", International Conference on Image Information Processing, IEEE, pp. 1-4. (2011).

[6] P.Singh and S.Budhiraja, "Feature Extraction and Classification Techniques in O.C.R. Systems for Handwritten Gurmukhi Script – A Survey" International Journal of Engineering Research and Applications (IJERA), Vol.1, pp. 1736-1739, (2012).

[7] J. Hussain and Lalthlamuana, "Artificial neural network-based approach for Mizo character recognition system", Science Vision, Vol. 14, No. 2, pp. 61-66,(2014).

[8] N.Kalchbrenner, E.Grefenstette and P.Blunsom," A Convolution Neural Network for Modelling Sentences", Annual Meeting of the Association for Computational Linguistics, pp. 655–665, (2014**).**