Pak. J. Statist.
2018 Vol. 34(1), 47-53

STATISTICAL FEATURES EXTRACTION FOR CHARACTER RECOGNITION USING RECURRENT NEURAL NETWORK

Saeeda Naz[1,4], Arif I. Umar[1], Saad B. Ahmed[2], Riaz Ahmad[3], Syed H. Shirazi[1], Muhammad I. Razzak[2] and **Amir Zaman[5]**

[1] Department of Information Technology, Hazara University, Mansehra, Pakistan. Email: saeedanaz292@gmail.com
arifiqbalumar@yahoo.com

[2] University of Technology, Kaiserslautern, Germany

[3] King Saud bin Abdulaziz University for Health Sciences, Riyadh, Saudi Arabia

[4] GGPGC, No. 1, Abbottabad, Higher Education Department, KPK, Pakistan

[5] Department of Education, Abdul Wali Khan University Mardan, Pakistan

ABSTRACT

Recent studies show that recurrent neural network provided promising results for character recognition. We have extracted number of features using sliding window approach from normalized Urdu Nasta'liq text line image. The text line is scanned from right to left and top to bottom by considering Urdu script properties and extracted geometrical or statistical features, zoning and raw pixels features. We conduct four studies like sliding window with non-overlapped frame, sliding window with overlapped area with previous frame, multiple zones in a frame and raw pixels. In this paper, we evaluated MDSLTM with CTC output layer on UPTI dataset for Urdu character recognition.

KEYWORDS

Recurrent Neural Network; Urdu OCR; Features Extraction.

1. INTRODUCTION

Recent advancement in pattern recognition algorithm made possible to develop robust character recognition system having similar properties as of human. Latin script character recognition has been extensively studied in the last half century and achieved its maturity level and delivered technology driven applications in the market. The text recognition systems have been developed to recognize Hangul characters, Japanese Kanji Chinese characters, Japanese Katakana syllabic characters, Chinese characters, etc., but it is not the same case for cursive languages like Arabic, Urdu, Farsi, Jawi, Pashtu and Sindhi etc. It may be due to the complexity of this script (See Naz et al., 2013, Naz et al., 2016 and Naz et al., 2107).

Optical Character Recognition (OCR) has number of steps like image acquisition, pre-processing, segmentation, features extraction, classification and post-processing. The feature extraction step is a crucial step for character recognition. This step is totally dependent on the problem directly and extracts the similar features in same class from the image of word or text line. The selection of features are depending on the type of OCR like online, printer or handwritten, type of script, nature of writing text, and type of segmentation approach like holistic or analytical. For selection of good features, some criteria need to be considered like easy computation, easy adaptation to other language, one feature do not replicate other feature and invariant especially in case of word or sub-word in cursive script. There are two main categories of features are structural and statistical. *Structural or topological features* extract from the topology or structure of the characters or sub-word. Some important structural features are curves, loops, joining points, horizontal line, vertical lines, diagonal lines, character's height and width, character's area and perimeter, etc. *Statistical or geometrical features* extract from arrangement of pixels in the image. Computations of these features are easy as compare to structural features. Some examples of statistical features are zoning, projection histogram, moments, mean, variance projection histogram, foregrounds etc.

In this paper, we are considering geometric features due to machine learning classifiers as statistical model works well on geometric features and can easily apply on other language also. Recent work for character recognition showed promising result for not only the character recognition of Latin script [See Nishide et al. (2011), Graves et al. (2009), Graves (2012), Graves and Schmidhuber (2009), Liwicki et al. (2007) and Graves (2013)] but also provide very good accuracy for Urdu script based languages [See Ul-Hasan et al. (2013) and Ahmed (2015)]. We presented *n* set of features extracted from overlapped sliding window from right to left and top to bottom. Section 2 presents the detail of set of features extracted for Urdu script character recognition using non-overlap-sliding window, overlap-sliding window and zoning approaches.

2. FEATURES EXTRACTION

In this article, we propose to extract number of features using (overlap) sliding window approach from normalized Urdu Nasta'liq text line image. After normalization, the height of the text line is fixed and the width is variable according to the length of text lines. Using sliding window approach, the text line scan from right to left and top to bottom by considering Urdu Nasta'liq language properties into number of frames and computed the geometric/statistical features from each frame/window. The width (w) of the window or frame (f) can be between one and ten. The number (n) of features are extracted from each frame of each text lines and appended in the feature vectors. So the Urdu text line is converted into n width vectors for the classifier. The following is the detail of extracted features.

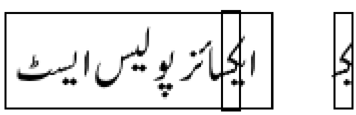

Fig. 1: Frame Extracted using Sliding Window from Urdu Nasta'liq Text Line

2.1 Horizontal Edges

For edges detection feature, sobel method applied to compute two dimensional gradient magnitudes at each point in each frame of the text lines to detect edges along row wise [See Khorsheed (2007)]. Then the total numbers of intensities of extracted horizontal edges are counted and append to the feature vector.

Fig. 2: Horizontal Edges in the Frame using Sliding Window

2.2 Vertical Edges

Using sobel method, the edges are calculated vertically and then count all the intensity values and concatenate to the feature vector.

Fig. 3: Vertical Edges in the Frame using Sliding Window

2.3 Foreground INTENSITY DISTRIBUTION

In the foreground distribution, the total pixels are counted and summed in the frame of text lines.

$$f_3 = \sum_{i,j}^{mn} p(i,j) \;\; if\; p(i,j) > 127 \tag{1}$$

2.4 Density Function

It is used to find out the ratio of summation of pixels in the foreground in each frame divide by total size of frame and contented to the feature vector of the txt line.

$$f_4 = \sum_{i,j}^{mn} p(i,j) \Big/ f_{size} \tag{2}$$

2.5 Mean and Variance of Horizontal Projection

The horizontal projections calculated using summation of pixels intensities in each row in a frame and then mean value is calculated of each frame and concatenate to the feature vector for each frame. Variance is also computed of the summed vertical projection in each frame of the text lines.

2.6 Mean and Variance of Vertical Projection

The vertical projections by summing of pixel intensities column wise in a frame of text line image and then mean value is calculated of vertical projection and concatenate to the feature vector for each frame. Variance computed of the summed vertical projection in each frame of the text lines.

2.7 Pixels Distribution in a Frame

The total numbers of pixels are summed from the frame of text line image and append to the feature vectors.

$$f_{10} = \sum_{i,j}^{mn} p(i,j) \tag{3}$$

2.8 Contrast and Energy

In the window frame, the contrast of intensity has measured between a current pixel and its neighboring pixels as in eq. 3. In feature 11, the gray level intensities are squared and then summed which measured the closeness or uniformity among the pixels distributions.

$$f_{11} = \sum_{i,j}^{mn} |i-j|^2 \, p(i,j) \tag{4}$$

$$f_{12} = \sum_{i,j}^{mn} p(i,j)^2 \tag{5}$$

2.9 Homogeneity and Correlation

In this feature, the angular second moment is calculated form each frame from start of the text line to end. And finally appended to the feature vector for classifier, in eq. 5. The correlation of the neighbors pixels are calculated in the frame of the text line image in eq. 6.

$$f_{13} = \sum_{i,j}^{mn} \frac{p(i,j)}{1+|i-l|} \tag{6}$$

$$f_{13} = \sum_{i,j}^{mn} \frac{(i-\mu i)(j-\mu j)p(i,j)}{\sigma_i \sigma_j} \tag{7}$$

2.10 Center of Gravity in x and y Direction

The first element of centroid has calculated from the x-coordinate. It gives the center if mass of the character or stroke in the frame of sliding window from text lines. In this feature, the sum of pixels intensities is calculated for foreground and also count the total number of foreground values. If the foreground intensities are greater than 0 then centroid of rows will be the ration of summation of rows divided by total number of pixels of foreground of character or stroke in each frame of the text line image otherwise centroid of row will be zero.

The second element of centroid has calculated from the y-coordinate. It gives the center if mass of the character or stroke in the frame of sliding window from text lines as explain above.

2.11 Zoning

In the study of zones feature extraction, each frame (f_i) content is separated from text line and divided into number of $n \times n$ zones (n_i) where $n = 3, 5, 7$. Different number of experiments performed to extract features from the zones of each frame of the text lines image. The height of frame (f_h) is dividing by number of cells [See Bunke (2004)]. The intensity values are summed from the foreground of the character in each zone (n_i of the frame and then divided by the total number of pixels values in all cells $(\sum n_i)$ in each frames as in eq. 8 and also divided by the total size of the frame (f_{size}) as in eq. 9. This feature extraction depends on the local averaging and more robust to distortion, noise and cursive writing property of the Urdu Nasta'liq.

$$f_{16} = \frac{n_i}{\sum n_i} \tag{8}$$

$$f_{17} = \frac{n_i}{f_{size}} \tag{9}$$

2.12 Pixels

Image is combination of pixels which has different intensities. These intensities have information of objects, and background. In this study, feature vector is populated by each pixel value of the image and then pass to the classifier for classification.

3. DISCUSSION

Recent research on feature extraction and selection for the character recognition shows the importance of feature extraction phase for character recognition. Researchers have used different features for the recognition of complex Urdu language. Each character is different due to different features and the character is recognized on the basis of these distinct features. We have considered statistical or geometric features due to machine learning classifiers; as statistical model works well on statistical/geometric features and can easily apply on other language also. In one study, we have extracted a set of *n* features with sliding window that slides from right to left and top to bottom. In second study, we have also moved the non-overlapped sliding window and fifty percent overlapped sliding window with window size of 2 pixels (1 columns overlap), 4 pixels (2 columns overlap), 6 pixels (3 columns overlap), and 8 pixels (4 columns overlap); and extracted features form each window. In last study, we divided the sliding window into number of zones to extract the local features of the text lines. The size of zones are 3×3, 5×5 and 7×7.

We test the extracted features using MDLSTM variant of RNN and evaluated that (overlapped) sliding windows based statistical features approach; zoning based approach or pixel based MDLSTM based approach. We also compared that which approach provide considerable gain in accuracy and overcome computational time complexity. We explored and concluded that manual features based approach overcome the computational time complexity issue while has low accuracy as compare to pixels based features. The pixel features based MDLSTM approach has high computational time but it has high accuracy. Four MDLSTM network models are trained and different accuracies achieved using different features vectors by different models as shown in Table 1.

Table 1
Accuracies Achieved by using Different Features using Implicit Segmentation and MDLSMT on UPTI Dataset

Proposed Systems	Features	Classifier	Recognition Rate (%)
System 1 (zones)	Zoning	MDLSTM	93.38
System 2 (slidding window)	Statistical	MDLSTM	94.97
System 3 (non-overlapped sliding window)	Statistical	MDLSTM	96.40
System 4 (raw pixels)	Raw pixels	MDLSTM	98

4. CONCLUSION

In this article, we extracted number of statistical features which measured the local characteristics of the text line image and also extracted raw features. The extracted features were encodes in feature vector from n-pixels width frame which was separated using (overlap) sliding window approach and zoning approach from Urdu Nasta'liq text line image. Finally, the feature vector set injected into character recognition engine. The character recognition engine is built using MDLSTM variant of Recurrent Neural Network (RNN) and CTC output layer. The RNNLIB is an open source speech precognition library. The proposed engine trained, validated and tested on the Urdu Printed Text Image (UPTI) database for comparing the results with the state of the art system.

REFERENCES

1. Naz, S., Hayat, K., Razzak, M.I., Anwar, M.W., Madani, S.A. and Khan, S.U. (2014). The optical character recognition of Urdu-like cursive scripts. *Pattern Recognition*, 47(3), 1229-1248.
2. Ahmed, S.B., Naz, S., Razzak, M.I., Rashid, S.F., Afzal, M.Z. and Breuel, T.M. (2016). Evaluation of cursive and non-cursive scripts using recurrent neural networks. *Neural Computing and Applications*, 27(3), 603-613.
3. Naz, S., Umar, A.I., Ahmad, R., Ahmed, S.B., Shirazi, S.H. and Razzak, M.I. (2017). Urdu Nasta'liq text recognition system based on multi-dimensional recurrent neural network and statistical features, *Neural Computing and Application*, 28(2), 219-23.

4. Nishide, S., Okuno, H.G., Ogata, T. and Tani, J. (2011). Handwriting prediction based character recognition using recurrent neural network. In 2011 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2549-2554.
5. Graves, A., Liwicki, M., Fernández, S., Bertolami, R., Bunke, H. and Schmidhuber, J. (2009). A novel connectionist system for unconstrained handwriting recognition. *IEEE transactions on pattern analysis and machine intelligence*, 31(5), 855-868.
6. Graves, A. (2012). Offline arabic handwriting recognition with multidimensional recurrent neural networks. In *Guide to OCR for Arabic scripts* (pp, 297-313). Springer London.
7. Graves, A. and Schmidhuber, J. (2009). Offline handwriting recognition with multidimensional recurrent neural networks. In *Advances in Neural Information Processing Systems*, 545-552.
8. Liwicki, M., Graves, A., Bunke, H. and Schmidhuber, J. (2007). A novel approach to on-line handwriting recognition based on bidirectional long short-term memory networks. In Proc. *9th Int. Conf. on Document Analysis and Recognition*, 1, 367-371.
9. Graves, A., Mohamed, A.R. and Hinton, G. (2013). Speech recognition with deep recurrent neural networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP), 6645-6649.
10. Ul-Hasan, A., Ahmed, S.B., Rashid, F., Shafait, F. and Breuel, T.M. (2013). Offline printed Urdu Nastaleeq script recognition with bidirectional LSTM networks. In *12th International Conference on Document Analysis and Recognition* (ICDAR), 1061-1065.
11. Ahmed, S.B., Naz, S., Swati, S., Razzak, M.I., Khan, A.A., Umar, A.I. (2015). UCOM offline dataset: A Urdu handwritten dataset generation. *International Arab Journal of Information Technology*, 14(2), http://ccis2k.org/iajit/PDF/Vol%2014,%20No.%202/8721.pdf
12. Khorsheed, M.S. (2007). Offline recognition of omni font Arabic text using the HMM ToolKit (HTK). *Pattern Recognition Letters,* 28(12), 1563-1571.
13. Bunke, H., Bengio, S. and Vinciarelli, A. (2004). Offline recognition of unconstrained handwritten texts using HMMs and statistical language models. *IEEE transactions on Pattern analysis and Machine intelligence*, 26(6), 709-720.