IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-12, Dec 2019 E-ISSN: 2347-2693

Multilingual-Word-Script Classification in Text Video Frames

Sunil C.[1*], Chethan H.K.[2], Raghunandan K.S.[3], G. Hemantha Kumar[4]

[1,2]Dept. of Computer Science and Engineering, Maharaja Research Foundation, Maharaja Institute of Technology, Mysore, Karnataka, India
[3,4]Dept. of Studies in Computer Science, Manasagangothri, University of Mysore, Mysore, Karnataka, India

[*]*Corresponding Author: sunilchaluvaiah87@gmail.com, Tel.: +91 9739459309*

DOI: https://doi.org/10.26438/ijcse/v7i12.8792 | Available online at: www.ijcseonline.org

Accepted: 20/Dec/2019, Published: 31/Dec/2019

***Abstract*—** Nowadays, achieving good results for the text classification of the multilingual scripts in arbitrary images in the videos is the most challenging task for the researchers. Most of the people depends on the internet and the digital world that makes difficult task to understand the multilingual script in various domain. Motivated from this, we proposed a text classification model for multilingual-word-scripts in video frames extracted from the videos which contains South Indian Multilingual Scripts namely, English, Tamil, Kannada, Malayalam and Telugu. Six-layer convolution neural network model has been used to classify the text to their respective classes. In this work we have castoff 600 word images from each script and total of 3000 word images that is extracted as the word images from the video frames for our experimentation. Our proposed model is proficient in accomplishing decent classification results when compared to existing conventional methods such as KNN and SVM classifier.

***Keywords*—** Deep Neural Networks, Text Classification, Multilingual Scripts.

I. INTRODUCTION

Nowadays, with invent of advanced technology and rapid development of digital and smart city, whereas implementation and utilization of the road cameras for numerous applications, that produces huge amount of data every day [1]. This leads to enormous collection of videos and images which contains multiple scripts. This has been created a huge challenge for all the researchers in the area of image processing and pattern recognition to classify the multilingual scripts in video and images. Therefore, automatic classification of text in videos has become and essential task for analysis of the videos. Hence, it is essential to improve a classification system for multilingual-word-scripts in video frames for an effective recognition [2].

From recent survey we can observe that multilingual-word-script image classification has not been carried out much by the researchers and so it is still an active area for research as it poses some challenging task for classifying multilingual-word-script into to their respective classes in video frames and some of them are complex background, blurriness, orientations, noise and low resolution. Some of the samples of the multilingual-word-script images that is extracted from the video frames are shown in the Fig. 1. In Fig.1. We can see the samples of multilingual-word-script images of (a) Kannada, (b) Malayalam, (c) Tamil, (d) Telugu and (e) English which is comprised of various challenges [3].

Figure 1. Shows the samples of multilingual word script images of (a) Kannada, (b) Malayalam, (c) Tamil, (d) Telugu and (e) English.

The work on the video frames for the classification of text have not been carried out much from the recent survey according to the best of our knowledge, especially based on

words. We can see some of the work done based on text line, word level and block level in document images [4] and some of the works are limited to two scripts. Nowadays, we can observe that the deep learning model have been achieved great success for the task of text classification as it is capable of learning text representation from the original data [5]. Considering all these factors we are motivated to propose multilingual-word-script classification of text with deep convolution neural network in video frames which are extracted from videos.

Sequentially, the proposed work of the paper is divided into sections such as, section 2 gives brief explanation of the literature survey, section 3 explains the proposed methodology, section 4 illustrates the experimentation and results, and section gives the conclusions and predicts the future scope.

II. RELATED WORK

A K Bhunia et. al., [6] proposed a novel method for identifying the text in video frames and natural scene images by extracting the local and global features based on convolution neural network with LSTM framework and weighting dynamically. Then by applying softmax layer attention-based patch weights are calculated. At last they use fusion method which is capable of forming individual patch from local and global features and achieved better results when compared to conventional methods. W Li et, al., [7] proposed a new technique that is built on the integration of the deep neural network and sentiment linguistic knowledge known as sentiment–feature-enhanced deep neural network (SDNN) for classification of text sentiment. For illustration of words they have integrated sentiment lexicon and attention mechanism that companions the gap between sentiment linguistic, conventional and deep neural network systems. The suggested work SDNN technique have achieved enhanced results after relating to state-of-the-art methods. M Z Amin et, al., [8] proposed a new text classification system for Question Answering Systems that is based on open domain. This model is built on convolution neural network classifier and classification model uses multiclass text classifier. For calculating the loss and mapping the semantically related words softmax layer is applied and gives better results for large scale text classification. M Hughes et, al., [9] proposed a new technique for classification of the medical text by constructing the convolution neural network. In this model sentence level classification system is built on medical documents and proved that use of the CNN for signify semantics of clinical text which permits semantic classification at sentence level. By making use of multi-layer convolution deep networks, it has a potential to produce more optimal features when compared to shallow learning methods. A Hassan et, al., [10] proposed new method for text classification on deep learning built on recurrent and convolution layers. This presented work is built on convolution neural network and bidirectional recurrent neural network based on bag of words. Instead of pooling layers they employ bidirectional layers to reduce loss in local information and to detention long-term dependencies in inputs and achieves better results when compared to sentiment analysis. K S Raghunandan et, al., [11] proposed a novel method for classification of images based on word type is different data types such as mobile camera, natural scene, video-scene and caption images. This model is based the integration of sharpness and contrast features and exploring the intensity and saturation spaces of HSI for classification of images at word level. For obtaining smoothed images and sharpen the edge details they have applied Maximum Value Difference operation. To obtain the feature vectors k-means clustering is used and SVM classifier for classification and achieved better results when compared to existing methods in terms of classification rate. J Mei et, al., [12] proposed a new method for identification scene text script based on deep convolutional and recurrent neural networks. This method utilize the image representation and spatial dependencies between the text lines and integrate the convolutional and recurrent neural network into one end-to-end trainable network, whereas, convolution network gives rich image representations and recurrent network successfully investigates long term spatial dependencies. To maintain the arbitrary sizes of the input images we adopt average pooling structure and performs excellent results when compared existing methods. S Roy et, al., [13] proposed a novel method for identifying of tampered information and classification of scene and caption text in video frames. This model introduce new method for classifying caption and scene texts by exploring the spatial distribution of DCT coefficients and adopts the unique way within the zero and non-zero coefficients of scene and caption texts to distinguish them and results proved to be classified effectively. N Sharma et. al., [14] proposed a novel method for classification of text frames in videos. The idea of this model is to categorize the text and non-text frames from input video frames built on linearity and non-linearity text components. For identification of text components they have integrated the color and gradient information from the RGB images of an input video frames. At if the linearity conditions is fulfilled from the components then it is considered as text or non-text components and this method is also compared with other existing methods and produces better results. P P Yeotikar et. al., [15] proposed a novel method for identification and separations of the text based on words into their respective languages such as Kannada, English, Hindi and also English numerals in tri-lingual documents. The proposed method is tests on the manually created data sets and produces good results and when test on the scanned documents the performance is not to that extent because it contains noise and skew errors. D Duong et. al., [16] proposed a novel method for classification of the sports

video based on the words. In this model they extract the SURF descriptors from the each key frames and by making use of k-means clustering technique they form visual words vocabulary. Histogram of these visual words are calculated and deliberated as the feature vectors and support vector machine is used to train each classifier to their respective classes and this bag of words model superior performance in the sort of sports video classification. S Haboubi et. al., [17] proposed different mode for segregating Arabic and Latin scripts and categorizing the Arabic words from these in the bilingual printed documents. For the separation of the words from Arabic and Latin they have adopted statistical and geometrical analysis and the words are extracted on the basis of the structural features and for classification the neural networks are applied and produces promising results. P Shivakumara et. al., [18] proposed a novel method for classification of the text frames in videos that is built on mutual nearest based symmetry. In this model for selection of the text blocks they have integrated wavelet with median moment with k-means clustering technique and the prevailing text pixels are recognized using the max-min clustering at the pixel level and mutual nearest neighbour based symmetry is used to obtain the text pixels and proves that text frame classification is necessary before text detection. S Chanda et, al., [19] proposed two-way method for identification of scripts such as English, Bengali and Devnagari from the printed documents based on word-wise model. In the first approach identification of scripts from noisy data, 64 dimensional chain-code histogram features are used and second approach for low resolution images with gradient features of 400 dimensional are adopted. Classification of the each script is based on the majority of voting of the character components and support vector machine method and achieves higher classification rate. W Zhang et, al., [20] proposed an novel method for classification of the text based on the representation of the multi-words. This multi-word extraction is done by using syntactical structure from these extracted multi-words two behaviours like concept and subtopic representation is presented to signify the documents. For classification the linear and non-linear in SVM are used and proposed method performs better when compared to general concept representation and linear kernel is better than non-linear kernel for classification. P B Pati et, al., [21] proposed a novel method for identification of multi-scripts based on words from the documents of the images. In this mode they have considered bi-script, tri-scrip and elven-script states and for evaluation of features of Gabor and Discreet Cosine Transform by making use of nearest neighbour and linear discriminant. For classification SVM classifiers are used and integration of these achieves better results. S Jaeger et, al., [22] proposed a new method for identification of the scripts based on the words from the bilingual printed documents. This model uses multiple classifier system for identification of the scripts and Gabor filter investigation on word level for bifurcating the Latin and non-Latin words and the entire system consists of Gaussian mixture models, weighted Euclidean Distances, nearest neighbours, and SVM for the languages such as Chinese, Hindi, Arabic and Korean and achieves good classification rate. A S Banu et. al., [23] proposed classification model for Synthetic Aperture radar (SAR) image based on wavelet transform and Euclidean distance by making us of shanon index measurement that involves three steps pre-processing , feature extraction and classification process. For extracting features they have used Daubechies wavelet and for classifying Euclidean distance by making use of shanon index measurement and proposed method is compared with other existing methods and achieved better accuracy. N S Lele., [24] proposed a classification of images based on convolution neural networks. This work is done on supervised learning as well as unsupervised image classification based on the CIFAR-10 dataset images by drawing rounded boxes around various images and by naming those images and proves to better when compared to traditional image classification algorithms.

III. PROPOSED METHODOLOGY

Multilingual-word-script classification of text in video frames have been created more awareness to computer vision techniques from past decades. From recent survey we can see that Convolution Neural Network (CNN) performs better when compared to other conventional methods because of its self-learning ability which can perform on various types of complex and huge data. This motivated us to utilize the Deep CNN for classification of text in video frames. In this work, we have adopted deep learning model that is 6 convolution neural network layers [25]. The 6 layers of the convolution neural network are shown in Table 1. In this process we carry out three phases and first phase is pre-processing by cropping the images that contains equal height and width, therefore the dimensions of cropped image should be smaller to the original image and finally we resize the all these images to a fixed size of 250x250x3 to set it as an input images in first layer of the deep learning network. The second phase is the feature transformation after the first layer it is followed by three fully connected layers to accomplish convolution and pooling processes. Each fully connected layers with two continuous convolution layers that is monitored by the max pooling layer. For activation all the layers in the model uses rectified linear units (ReLU), as it is consider as the modest non-linear function that can be adopted for activation processes. Rectified Linear Units is defined as [26] as shown in Eq. (1),

$$f(x) = \begin{cases} x, & x \geq 0 \\ 0, & x < 0 \end{cases} \quad (1)$$

Table 1. Parameter setting for CNN Input Size(In), Kernel(K), Stride(S), Padding(P) and Feature Maps(FM).

Type	Parameters
Input	250x250x3
Conv_1	In=250x250, K=3x3, S=1, P=1, FM=16
Batch_Norm	
ReLU	250x250
Max_pool	In=250x250, W=2x2, S=2
Conv_2	125x125, K=3x3, S=1, P=1, FM=32
Batch_Norm	
ReLU	125x125
Max_pool	In=125x125, W=2x2, S=2
Conv_3	62x62, K=3x3, S=1, P=1, FM=64
Batch _Norm	
ReLU	62x62
Max_pool	In=62x62, W=2x2, S=2
Conv_4	32x32, K=3x3, S=1, P=1, FM=128
Batch _Norm	
ReLU	32x32
Max_pool	In=32x32, W=2x2, S=2
Conv_5	16x16, K=3x3, S=1, P=1, FM=256
Batch_Norm	
ReLU	16x16
Max_pool	In=8x8, W=2x2, S=2
Conv_6	8x8, K=3x3, S=1, P=1, FM=256
Batch_Norm	
ReLU	8x8
Fully connected layer Input size	
Softmax prediction for labels	

In In this model for the convolution we have used the kernel size 3x3 and for max pooling 2x2 and the stacked layers represent the input image at different levels of abstraction. The third one is the classification whereas, all the 2 dimensional images are compressed to one dimension feature vector that was learned from the earlier steps and it is ended with the output layer. The fully connected layer of neurons that consists of the same number of classes is the output layer and by making use of the softmax it will output he probability for all individual class. At last we adopt the dropout layers as it regularize the network and protect from overfitting by adding them before the output layer and after each pooling layers.

IV. EXPERIMENTATION AND RESULTS

In this paper we propose multilingual-word-script classification of text in video frames. Each scripts are extracted as word images from the video frames. For illustration we have built the word scripts database that is extracted from the our own multilingual South Indian datasets of each 600 word images from the each scripts and total of 3000 word images form all the 5 scripts such as English, Kannada, Tamil, Telugu and Malayalam. For classification results of the proposed model is done obtaining confusion matrix. Table.1 shows the confusion matrix for deep CNN. We have compared our proposed model with KNN and SVM classifiers. For the classification using SVM we have used RBF kernel and for the calculation of the accuracy we have used 10 fold cross validation with several kernel parameters and cost parameters for each binary classifier. Table 2 shows the confusion matrix for KNN classifiers and Table 3 shows the confusion matrix for SVM classifier and from the classification results we can observe that deep CNN outperforms compared to other conventional classifiers such as KNN and SVM. Fig.2 and Fig.3 shows the successful and unsuccessfully multilingual-word-script classified images of the proposed method.

Table.2. Confusion matrix for deep CNN

Classes	English	Kannada	Tamil	Telugu	Malayalam
English	**95.6**	1.2	1.8	1.4	0
Kannada	3.8	**86.3**	4.7	2.9	2.3
Tamil	0	1.4	**88.4**	2.4	7.8
Telugu	0	6.8	2.4	**89.2**	1.6
Malayalam	1.7	8.5	4.3	2.7	**82.8**

Table.3. Confusion matrix for KNN Classifier

Classes	English	Kannada	Tamil	Telugu	Malayalam
English	**86.6**	3.9	2.6	3.8	3.1
Kannada	2.6	**83.8**	3.6	6.3	3.7
Tamil	2.4	4..7	**80.3**	4.5	8.1
Telugu	1.6	12.6	5.6	**77.4**	2.8
Malayalam	3.4	5.7	8.5	3.9	**78.5**

Table.4. Confusion matrix for SVM classifier

Classes	English	Kannada	Tamil	Telugu	Malayalam
English	**84.4**	4.2	3.9	4.5	3
Kannada	1.7	**82.3**	2.9	9.3	3.8
Tamil	2.5	5.3	**78.6**	3.4	10.2
Telugu	1.9	12.7	3.8	**75.3**	6.3
Malayalam	2.1	11.2	4.2	5.7	**76.8**

Figure 2. Sucessfully classified multilinugal-word-script images a.) Kannada, b.) Tamil, c.) Malayalam, d.) Telugu, e.)English.

Figure 3. Unsucessfully classified multilinugal-word-script images a.) Kannada, b.) Tamil, c.) Malayalam, d.) Telugu, e.)English.

V. CONCLUSION AND FUTURE SCOPE

In this paper, we propose a classification of the multilingual-word-script extracted from the video frames based on Deep Convolutional Neural Networks for Kannada, Tamil Malayalam, Telugu and English. This work is comprised of six-layer convolutional neural network model and max pooling. The proposed method produces superior performance when compared to other two conventional methods such as KNN and SVM classifiers. Future scope the work is to develop a Multilingual OCR for South Indian scripts based on this classification model.

ACKNOWLEDGMENT

The presented model in the above paper is also supported by High Performance Computing Lab, under UPE Grant Department of Studies in Computer Science, University of Mysore, and Mysore.

REFERENCES

[1] L Pang, S Zhu and C W Ngo., "Deep Multimodal Learning for Affective Analysis and Retrieval", IEEE Transactions, Multimedia, Vol. **17**, pp. **2008-2020, 2015.**

[2] M M Rathore, A Paul, A Ahmad and S Rho., "Urban planning and Building Smart Cities based on Internet Things using Big Data Analytics", Computer Networks, pp. **63-80, 2016.**

[3] P B Pati and A G Ramakrishana., "OCR in Indian Scripts: A Survey", Journal of IETE Technical Review, pp. **217-227, 2015**.

[4] D Ghosh, T Dube and A P Shivaprasad., "Script Recognition – Review", IEEE Transactions, pp. **2142-2161**, PAMI **2010.**

[5] T Young, D Hazarika and S Poria., "Recent Trends in Deep Learning based onNatural language processing", IEEE Computational Intelligence Magazine, Vol **13**, Issue **3**, pp **55-75, 2018.**

[6] A K Bhunia, A Konwer, A K Bhunia, A Bhowmick, P P Roy and U Pal., "Script Identification in natural scene image and video frames using attention based Convolutional-LSTM network", Pattern Recognition, Elsevier, Vol **85**, pp. **172-184, 2019.**

[7] W Li, P Liu, Q Zhang and W Liu., "An Improved Approach for Text Sentiment Classification Based on Deep Neural Network via a Sentiment Attention Mechanism", Journal of Future Internet, **11940, 2019.**

[8] M Z Amin and N Nadeem., "Convolution Neural Network: Text Classification Model for Open Domain Question Answering System", Computer Science, Information Retrieval, **2019.**

[9] M Hughes, I Li, S Kotoulas and T Suzumura., "Medical Text Classification Using Convolution Neural Networks", Studies in Health Technology and Informatics, Vol 235, pp. **246-250, 2017.**

[10] A Hassan and A Mahmood., "Efficient Deep Learning Model for Text Classification based on Recurrent and Convolutional Layers", 2017 16th IEEE International conference on Machine Learning and Applications (ICMLA), pp. **1108-1113, 2017.**

[11] K S Raghunandan, P Shivakumara, G H Kumar, U Pal, and T Lu., "Sharpness and Contrast Features for Word-Wise Video Type Classification", 2017 4trh IAPR Asian Conference on Pattern Recognition (ACPR) pp. **103-108, 2017.**

[12] J Mei, L Dai, B Shi and X Bai., "Scene Text Script Identification with Convolutional Recurrent Neural Networks", 2016 23rd International Conference on Pattern Recognition (ICPR), pp. **4053-4058, 2016.**

[13] S Roy, P Shivakumara, U Pal, T Lu and C L Tan., "New tampered features for Scene and Caption text Classification in Video Frames",2016 15th International Conference of Frontiers in Handwriting Recognition (ICFHR), pp. **36-41, 2016.**

[14] N Sharma, P Shivakumara, U Pal, M Blumenstein and C L tan., "Piece-wise Linearity based Method for Text Frame Classification in Video", Pattern Recognition, Elsevier, Vol **48**, pp. **862-881, 2015.**

[15] P P Yeotikar and P R Deshmukh., "Script Identification of Text Words from Multilingual Document", International Journal of Computer Applications, pp. **22-29**, X-PLORE **2013.**

[16] D Duong, T Ba Dinh, T Dinh and D Duc., "Sports Video Classification using Bag of Words Model", Intelligent Information and database Systems, ACIIDS, Springer, Vol. **7198**, pp. **316-325,2012.**

[17] S Haboubi, S Maddouri and H Amiri., "Word Classification in Bilingual Printed Documents", 2012 6th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT), pp. **502-506, 2012.**

[18] P Shivakumara, A Dutta, T Q Phan, C L Tan and U Pal., "A Novel Mutual Nearest Neighbor based Symmetry for Text Frame Classification", Pattern Recognition, Elsevier, Vol **44**, Issue **8**, pp. **1671-1683, 2011.**

[19] S Chanda, S Pal, K Frankle and U Pal., "Two-stage Approach for word-wise script Identification", 2009 10th International Conference on Document Analysis and Recognition, pp. **926-930, 2009.**

[20] W Zhang, T Yoshida and X Tang., "Text Classification based on multi-word with Support Vector Machine ", Knowledge Based Systems, Elsevier, Vol **21**, Issue **8**, pp. **879-886, 2008.**

[21] P B Pati and A G Ramakrishana., "Word Level Muli-Script Identification", Pttern Recognition Letters, Vol **29**, Issue **9**, pp. **1218-1229, 2008.**

[22] S Jaeger, H Ma and D Doermann., " Identifying Script on Word-Level with Informational Confidence", Eighth International Conference on Document Analysis and Recognition (ICDAR'05), Vol **1**, pp. **416-420, 2005**.

[23] A S Banu, P Vasuki, S M M Roomi and A Y Khan., "Sar Image Classification by Wavelet Transform and Euclidean Distance with

Shanon Index Measurment", International Journal of Scientific Research in Network Security and Communictions (IJSRNSC), Vol **6,** Issue **3,** pp **13-17, 2018.**

[24] N S Lele., "Image Classification using Convolution Neural Network", International Journal of Scientific Research inComputer Science and Engineering (IJSRCSE), Vol **6,** Issue **3,** pp **22-26, 2018.**

[25] M S Hossain, M Al-Hammadi and G Muhammad., "Automatic fruit Classification using Deep Learning for Industrial applications" In IEEE Transactions on Industrial Informatics, **2015**, pp.**1027-1034.**

[26] G E Dahl, T N Sainath and G E Hinton., "Improving Deep Neural Networks for LVSCR using Rectified Linear Units and Dropout", 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, **2013**, pp. **8609-8613.**

Authors Profile

Mr.Sunil C received degree in Bachelor of Electrical and Electronics Engineering and M.Tech in Bio-medical Signal processing and Instrumentation from Visvesvaraya Technological University, Belgaum, Karnataka. Currently he is pursuing Ph.D. at University of Mysore, Karnataka. His research interest includes image processing, pattern recognition, video understanding and Bio-medical Image Processing.

Mr.Raghunandan K S received masters from University of Mysore in the Year of 2013. Currently, he is pursuing Ph.D. at University of Mysore, Karnataka, India. His research interest includes image processing, pattern recognition and video understanding. He has published many papers in International Conferences and Journals.

Dr.Chethan H K received Bachelor's, Master's and Doctorate degree from University of Mysore, Karnataka, India. Currently working as Professor at Maharaja Institute of Technology, Thandavapura, Karnataka India. Guiding 8 Ph.d Students in several domains. Have guided several projects for bachelors and masters' student. He has published many papers in International conferences and Journals.

G. Hemantha Kumar received B.Sc., M.Sc. and Ph.D. fromUniversity of Mysore. He is working as a Professor in the Department of Studies in Computer Science, University of Mysore, Mysore. He has published more than 200 papers in Journals, Edited Books and Refereed Conferences. His current research interest includes Numerical Techniques, Digital Image Processing, Pattern Recognition and Multimodal Biometrics.

