International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-4, November 2019

Deep Features based Multilingual Text Detection and Recognition in Video Frames

Sunil C, K S Raghunandan, H K Chethan, G Hemantha Kumar

Abstract*: One of the major causes for achieving poor text detection results in video frames are complex background, illumination and de-blurring of the frames and it is also important challenges for the researchers to overcome such problems. Therefore, motivated from this kind of observation from recent survey, we propose a text detection method based on Deep Neural Networks known as TextBoxes which is capable of detecting text in video frames with improved performance when compared to state-of-the-art techniques. In parallel this we also propose a Text candidate detection for video frames and scene images by extracting words based on Automatic Window Detection by making use of Discrete Wavelet Transform (DWT) with the sliding window for extracting high frequency sub bands for each sliding window. K-means clustering technique has been used to obtain the text components and to decrease the background complexity and noise. Six-layer convolutional neural network model has been designed to recognize the text in multilingual images. Experiments for text detection are done on our own multilingual South Indian database, ICDAR-2015 Videos, YVT videos, SVT, and MSRA Scene datasets and demonstrated in terms of Recall, Precision and F-measure and for recognition ICDAR-2015 Videos, ICDAR 2011 and SVT scene images.

Keywords: Text Detection, Text Recognition, Deep Neural Network.

I. INTRODUCTION

Optical Character Recognition is normally abbreviated as OCR and it has been the most important area of research in the field of pattern recognition. The basic idea of OCR is to recognize the characters present in the images and videos without any intervention of humans. The excessive usage of smart phones and increasing growth of the internet and communication through social networks as caused to produce huge amount of visual data. Hence, videos are streaming in online because of uploading and downloading of videos in massive form. This huge data has triggered many researcher for understanding the multimedia and information retrieval form videos. Thus, it has created a challenging task for detection and recognition these text content present in the videos [1].

Normally text in videos can be classified as scene text and artificial text. For example text in news videos in the form of captions gives you the information about the events and in speech videos the subtitles provide us a basic information. However we concentrate on detecting and recognizing the text from video frames that is extracted from the videos in the proposed work. Usually video OCR method incorporates text detection, localization, extraction and recognition whereas, text regions are identified during the stage of text detection and to identify the exact area of text lines localization is adopted, text are extracted from complex background at the stage of extraction and recognition of the text are done using external OCR software's. Hence extracting and recognizing text successfully from multimedia information leads to some potential applications such as video indexing and retrieval, for assisting tourist, which helps to visually impaired persons and automation in supermarkets [2].

Text detection is first step in the process of extracting textual information from videos and from past few years conventional methods [3,4] and deep neural network based detection have been improved significantly for large datasets. Recent advances based on deep ConvNets proposed novel methods and achieved outstanding results on standard benchmarks [5,6,7,8] and region based convolution neural network technique(R-CNN) for object detection have achieved better performance by making using of deep neural networks [9], without sharing computation while novel method spatial pyramid pooling networks (SPPnets) have been adopted for increasing the speed of R-CNN with sharing computation [10].

Subsequently, the proposed work is organized as, section 2 address the previous work that is related to detection and recognition of text in videos and section 3 gives the explains the proposed methodology text detection and recognition from video frames, section 4 gives the experimentation and results of the proposed work and section 5 concludes and predict future scope.

II. RELATED WORK

K M Elizaveta et. al., [11] proposed a novel method for clustering of text in documents as graph community detection. In this proposed method the clustering of text does not make use of user set of parameters, hence the text in the documents and which relates between them is constituted as graph nodes and edges.

Manuscript published on November 30, 2019.
* Correspondence Author
Sunil C*, Department of Computer Science and Engineering, MIT Mysore, India. Email: sunilchaluvaiah87@gmail.com
K S Raghunandan, Dept. of Studies in Computer Science, University of Mysore, Mysore, India. Email: raghu0770@gmail.com
H K Chethan, Department of Computer Science and Engineering, MIT Mysore, India. Email: hkchethan@gmail.com
G Hemantha Kumar, Dept. of Studies in Computer Science, University of Mysore, Mysore, India. Email: ghk.2007@yahoo.com

Finally the clustering of text is done by making use of graph community detection technique and achieved good results. L Shangbang et. al., [12] proposed a novel method for detection of text in arbitrary forms by flexible text representation known as TextSnake. Horizontal, curved and oriented shape scene text presented in real world scenarios can be successfully represented through TextSnake and fully convolution network is adopted evaluate geometry attributes. Our proposed method TextSnake gives better results when compared to other state-of-art methods. Y Nagaoka et. al., [13] proposed a novel method for detection of text using multiple RPN (region proposal networks) with help of original faster R-CNN. Usually in original faster R-CNN the ROI (Region of Interest) which ids generated by region of proposal network with feature map from last convolution layer, but in this novel method multiple ROI are generated with multiple region proposal network by making use of feature map with multiple convolutional layers. ROI merge layer is also introduced to select correct ROI with effective multiple RPN. Major drawback is that it is limited to only horizontal text detection. L M Franics et. al., [14] proposed a novel method for text detection in various natural scene images with implementing least square Support Vector Machine (SVM) technique. The module developed is called as TEDLESS, the images are enhanced and then binary classification to classify text candidates from non-text candidates and then recognition of text and this is usually helpful for visually impaired people to identify the real world objects without any assistance of people but with help of technology. P Shivakumara et. al., [15] proposed a novel method for detection and recognition of text or bib number in marathon images by combining both torso and text detection. Features are extracted using histogram orientation gradients and then detection of upper body is done by SVM classifier and segmenting foreground from the upper body is done by grab cut technique. Hence this method has proved to achieve better results after torso detection then before torso detection. V N Manjunatha et. al., [16] proposed a novel model for identifying South Indian multi scripts from videos and images by adopting Gabor filter and k-means clustering procedure. Extraction of texture features and to sharpen the edges Gabor filter and wavelet transforms are combined and the output of the Gabor image is classified by k-means clustering technique. Morphological operations are applied for clustered pixels and for detecting the text line linked list approach is used and finally by using wavelet entropy the text regions are identified and experimentations shows the promising results. Y Wang et. al., [17] proposed a new model for detection of Chinese text in the news videos with multi-stage techniques. For identifying the text blocks Stroke width Transform with enhanced version is used. Divide and conquer procedure is used for discriminating the text from non-text blocks with design of Multi-Layer Perceptron networks and at last all the text blocks are integrated to form text lines. M V Toro et. al., [18] proposed a novel model for text detection of multi scripts and verification in road scene datasets. Maximum Stable Extremal Regions (MSER) along with Histogram of Stroke Width (HSW) is fused in this technique and all the characters will be combined together by raycasting to form a text lines and it is verified with HSW. Support Vector Machine (SVM) classifier is used for classifying the scripts into respective class and the model gives effective detection of the multiple scripts. A B Ayed et. al., [19] proposed novel technique which is based on parallel implementation by Map reduce for detecting texts in Big data natural scene videos. Detection of text is based on dividing the video frames into blocks and then each block is extracts the text and then it classified using Map reduce parallel programming technique. To verify the effectiveness of the model it is tested on YouTube video datasets and it shows that this method is twice faster than the classic methods. Lei Sun et. al., [20] proposed a novel and robust approach for automatic detection of text in natural scene images that is based on neural networks along with color-enhanced contrasting extremal regions (CER). According to this model the input images will be extracted as character candidates and also by classifying them into six component trees based on saturation, hue and gray scale. Later divide and conquer approach is adopted for labelling the each candidate image and neural networks are used for classification of text and non-text and method archives better performance on ICDAR datasets. X Wang et. al., [21] proposed a new model for text detection with implementation of higher order conditional random fields (CRF) and multi-layer segmentation from natural scene images. This model will segregate the text and background from the input image by using multi-layer segmentation technique and it forms 9 layers. Candidate text will be derived from these 9 layers with help of connected components (CCs) and later on it is verified by higher order CRF based analysis and this test on ICDAR datasets and proves to be improved. C Shi et. al., [22] proposed a new manner in which they have implemented Maximally Stable Extremal Regions (MSERs) technique for scene text detection in images. With MSERs they have made use of graph model based algorithm to minimize the cost function and other sources of information is put under one framework. To conclude, MSERs labeling procedure is used to eliminate the non-text regions and to detect the text region. Therefore, proposed system shows an improved results when related with other existing procedures in terms of precision and recall. V N Manjunath et. Al., [23] proposed a novel method for detection of multilingual scripts in both color and normal images. This model consists of two phases and in first phase they will combine the wavelet transforms and Gabor filter for extraction of sharp edges and textual features from the given input images. In second phase, for the resultant Gabor image the wavelet entropy is applied to know the energy value which in turn find out the high frequency texture elements of processed one to detect the region of text and detection accuracy is improved when compared to other existing method. P Shivkumara et. Al., [24] proposed novel method for text detection and classification of graphic and scene text in video images with complex backgrounds. This model make use of filters and edge based rules for the classification of high and low contrast video images for detecting text lines precisely.

The present model was compared with other three existing methods and proved to give better results. M Anthimopoulos et. al., [25] proposed a novel method for detecting text in video images based on two-stage system. In phase one the edge based method is used to detect text lines with high recall and in second phase, to reduce the false alarms the sliding window and SVM classifier are used and methods produce good results. Qixiang Ye et. al., [26] proposed a novel method for detecting text from video frames and in this with the help of multiscale wavelet features they intend a different method called coarse-to-fine procedure which is used to detect the text lines even in complex backgrounds. The coarse detection initially analyses the wavelet energy feature to recognize the possible text pixels and to join these pixels density based region growing method is adopted and by structural information the candidate text lines are separated and in fine detection by extracting four ways of texture features the text line is obtained and SVM classifier is used to classify the text. C W Lee et. al., [27] proposed a novel method of detecting text and removal from the video sequences by making use of SVM and spatiotemporal restoration. In this model the automatic detection of text is done using SVM based texture classifier and spatial and temporal restoration are joined to restore the background regions and methods proves to give better results.

A K Bhunia et. al., [28] proposed a novel method for recognition of text in video frames and scene images based on color channel selection by selecting automatically. From the selected color channel the features extracted based on hidden Markov Model and by making use of sliding window color channel selection process is done and SVM is used to classify the color channel which provides better text recognition. S Shetty et. al., [29] proposed a novel method for recognition and extraction text in video frames. In this method first the videos frames are extracted from the videos and then each frames are will be passed through the two phases one is text localization and another is text verification phase. Hence the result of this model is the detection of text from the video frame in form of word file. X Rong et. al., [30] proposed a novel method for recognition of scene text in video frames that also tracks the text region captured in a moving camera. Scene text recognition framework is represented for multiple frames and text detection is used for extracting the bounding boxes from the initial location of the first frame and till rest of the frames of the video. Then scene text character is used for character prediction and to configure text words conditional random field is applied and this combination of text detection and tracking have produced better accuracy in practical applications. W Ohyama et. al., [31] proposed a novel method for recognition of Arabic text in video frames that comprises of text-line detection, word segmentation and word recognition. In this model both text detection and recognition are integrated, where as true text-line are detected with higher recall rate and false alarm are diminished in word recognition method and hence integration of both detection and recognition archives better precision and f-measure for text detection. P Shivakumara et. al., [32] proposed a novel method for detection of text in videos captured from mobile cameras based on fractals and fractal expansion. Self-similarity is the property of fractals that is used in a different way for enhancement of mobile videos and for separating text and non-text k-means clustering is used. Hence to improve the text detection temporal frames make use of optimal flow estimation and archives good results for different fonts, multi-oriented, background variation and arbitrary texts. D Cheng et. al., [33] proposed a novel method for detection and recognition of text in videos frames. Localization of the text, normalization of text size and for background machine learning text verification are performed during text detection process. Text recognition is applied on text detected line by performing text segmentation and then followed by OCR method.

III. PROPOSED METHODOLOGY

In this work, we propose a text box detectors for detecting text in video frames and automatic window size detection for text recognition and the textbox layer architecture is the resemblance of VGG-16 architecture, keeping the layers from conv1_1through conv4_3 and by making use of parameter down-sampling the last 2 fully connected layers of VGG-16 is converted to convolution layers. Finally some more extra convolutional and pooling layers will be added. The textbox layers are also known as multiple output layers, which will be inserted at final and some intermediate convolutional layers and its outputs will be cumulative and endure a non-maximum suppression (NMS) procedure and all the output layers are also conservative. Only convolutional and pooling layers are involved in Text box layers and hence it is fully convolutional and it is familiar to arbitrary size images for testing and training. The methodology consists of three sub sections such as section A is text box layers for text detection, section B automatic window size prediction for text candidate detection and section deep feature neural network model for word recognition.

A. Text Box Layers for Text Detection

The major key component of the text-box layers are text boxes and based on the condition of feature map of input images textbox layers concurrently predicts presence of the text and bonding boxes [34]. The outputs during the map location are classification scores and offsets to its associated default boxes in convolutional manner. If we consider the image and feature map size correspondingly (p_{in}, q_{in}) and (p_{map}, q_{map}). The map location (i, j) which companions with default box $b_0 = (x_0, y_0, p_0, q_o)$ and hence the values of the text box layers are anticipated as $(\Delta x, \Delta y, \Delta p, \Delta q)$ indicating the box $b = (x, y, p, q)$ is detected with confidence c, where

$$\begin{aligned} x &= x_0 + p_0 \Delta x, \\ y &= y_0 + q_0 \Delta y, \\ p &= p_0 \exp(\Delta p), \\ q &= q_0 \exp(\Delta h). \end{aligned} \qquad (1)$$

According to the matching scheme presented in [35] we match the ground truth for the word boxes to the default boxes pursuant to box overlap during the training phase. Hence all the multiple default boxes with different sizes will be linked with specific map location and with the consideration of aspect ratio and scales the words will be divided virtually which helps to handle words with similar size by learning the textboxes with specific regression and classification weights. Usually when compared to general objects the words contains a huge aspect ratio and hence, building the textboxes is challenging task. To overcome this long default boxes will be adopted. In this method we have defined 1,2,3,5,7 and 10 total of 6 aspect ratios for the default boxes and in the horizontal direction it will be dense and in vertical direction it will be sparse and all the default boxes are set to the vertical offsets to overcome the issue. To avoid the noisy signals we adopt irregular 1x5 convolutional filters rather than standard 3x3 filters in the textbox layers and these filters produce rectangular respective fields.

For learning, by adopting the loss function from [35] we consider x to be the match indication matrix, ground truth g, predicted location l and confidence is c. Precisely the $i-th$ default box and the $j-th$ ground truth, $x_{ij}=1$ means matching while $x_{ij}=0$ otherwise. Definition of loss function is given by,

$$L(x,c,l,g)=\frac{1}{N}(L_{conf}(x,c)+\alpha L_{loc}(x,l,g)) \qquad (2)$$

Where,

N = number of default boxes which match the ground truth boxes,

α = is set to 1,

L_{loc} = smooth L1,

L_{conf} = 2-class softmax loss.

B. Automatic Window Size Prediction for Text Candidate Detection

Basically for recognition of the word first we should fix the window size for reach characters by making use of sliding widow. Usually most of the videos due to low quality it comprises of complex background and noise and to overwhelm this factors the best solution is to use high frequency of wavelet. So in this work an Automatic Window Detection for Text Recognition by making use of Discreet Wavelet Transform (DWT) has been used as presented in [36] and for each window the calculation of the orientation has been done by making use of Principal component Analysis (PCA). Consequently by applying DWT for each window image horizontal, diagonal and vertical high frequency sub-band images will be formed and various arbitrary word data images will always suffer from the noise, low resolution and it is very difficult to get informations of the edges from the characters. For that reason, such informations can be accomplished by high frequency sub-bands. For each of the three high frequency sub bands we apply k-means clustering technique with k=2 and for each high frequency sub bands it produces 2 clusters images. To identify the pixels of the text in the images we use the high cluster images from each sub bands and we take into consideration of horizontal and vertical sub band images by neglecting the diagonal sub band images. To get the fused sub band images we perform the OR operation on the horizontal and vertical sub bands and at last to find out the orientation of the each character of these fused sub band images we implement the principal component analysis and from these obtained fused images the pixels containing white are known as text components and pixels containing with black are known as background.

C. Word Recognition based on Deep Neural Network

As per the observation from past few years computer vision has extended a lot of attention towards the text detection and recognition and convolutional neural network has shown improved performance when compared to the conventional methods as it is capable to handle huge data with complex nature by the ability of self-learning. The information from the image patches can be extracted by knowing the size of the kernel with the help of convolutional neural networks. CNN illustrate a significant role in designing the kernel size of the each layer, whereas the kernel size is directly related to the information attained [37]. Typically, the local information is extracted from small kernel of the given patches and this is helpful to capture minute information. In this proposed work, the kernel size is fixed to 3x3 that is minimum for all the convolution layers to obtain the minute spatial information. Table.1 shows the six layer CNN architecture designed for character training model. For mapping the input features and for test of the sliding window we make use of the training images of 250x250 with same size. In the entire phase of training, a set of hyper parameters are designed with the segmented training images of the size 250x250 and according to the datasets to be utilized.

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

Type	Parameters
Input	250x250x3
Conv_1	Is=250x250, K=3x3, S=1, P=1, F_m=32
Batch_Norm	
ReLU	250x250
Max_pool	Is=250x250, W=2x2, S=2
Conv_2	125x125, K=3x3, S=1, P=1, F_m=64
Batch_Norm	
ReLU	125x125
Max_pool	Is=125x125, W=2x2, S=2
Conv_3	62x62, K=3x3, S=1, P=1, F_m=128
Batch _Norm	
ReLU	62x62
Max_pool	Is=62x62, W=2x2, S=2
Conv_4	32x32, K=3x3, S=1, P=1, F_m=256
Batch _Norm	
ReLU	32x32
Max_pool	Is=32x32, W=2x2, S=2
Conv_5	16x16, K=3x3, S=1, P=1, F_m=512
Batch_Norm	
ReLU	16x16
Max_pool	Is=8x8, W=2x2, S=2
Conv_6	8x8, K=3x3, S=1, P=1, F_m=1024

Batch_Norm	
ReLU	8x8
Fully_connected layer Input_size	
Soft_max prediction for class_labels	

Hyper parameters plays an important role in facilitating the process of learning in efficient and effective manner. For setting up an extensive experimentation we define some of the Hyper parameters like, padding, momentum, stride, learning rate, kernel size (ideally small), number of output from each layer, to learn a small batch of images, batch normalization (BN) factor, activation function Rectified Liner unit (ReLU) and pooling. The input images is convolved with the size of kernels 3x3 and the padding is done of the size 1x1 and it is defined as *(k-1)/2*, where 'k' is kernel size that is use to capture the local path information in the form of extracting weights. Mini batch BN layers are used to reduce the internal covariance between the set of all images. As the input images are small in size, it is ensured that with a small stride the loss of information will be negligible less. Hence the striding is fixed to one in our proposed model. To preserve the size of input image further we make use of non-linearity ReLU is added to output of the convolution layer. The fastness and the influence of the ReLU will leads to reduce the vanishing gradient issue and it also acts as element gradient function in the network and is given by Eq. (3).

$$f(x) = \begin{cases} x, & x \geq 0 \\ 0, & x < 0 \end{cases} \quad (3)$$

To improve the performance of classification, we have reduce the dimension of feature map, a max-pooling layer with a window size 2x2, stride and padding of one is fixed after the ReLU process. To train the CNN we have used Stochastic Gradient Decent Momentum (SGDM) by fixing learning rate to 10^{-1}. Huge number of data overfitting problem will occur and model does not gain stable after 20 epochs, in Fig. 1 shows how the learning rate is decreased and convergence rate will be achieved at 10^{-5} with 50 epochs. Eventually, for effective detection of the test input image softmax layer is used.

IV. EXPERIMENTATION AND RESULTS

In this paper we have proposed two contributions one is text detection and another one is text recognition. For evaluation of our proposed method we have considered our own multilingual datasets of 30 videos and standard benchmark datasets like, ICDAR-2015 of 23 videos and YVT of 30 videos and also SVT and MSRA scene datasets of 150 images each. Hence for evaluation of the capability of the multi-oriented text detection, we have consider the datasets that is composed of South Indian Multilingual scripts such as English, Kannada, Tamil, Malayalam and Telugu. Key frame selection method has been used to extract the frames from the videos and from each we have selected 5 key frames for evaluation of the text detection results and we have considered 115 key frames from the ICDAR-2015 videos, 150 key frames from YVT videos and 150 key frames from our own multilingual video datasets.

D. Text Detection Approach

D.1. Experiments on ICDAR 2015 and YVT Videos.

Experiments and results conducted by our proposed method on the ICDAR-2015 and YVT video datasets are shown in Fig.2 and Table II and III gives the results in terms of Recall, Precision and F- measure and it is also compared with other existing methods and we can notice the our proposed method illustrations the improved performance than existing methods in terms Recall, Precision and F-measure. Fig.1. shows the examples of text detection results of proposed method for ICDAR 2105 videos and YVT videos datasets. For the purpose of comparison with our proposed method we have consider other existing methods such as Li et al. [38], Liao et al. [39], Dey et al. [40], Yin et al. [41], Epshtein et al. [42] and Zhao et al. [43].

Table II. Evaluation of the proposed and existing methods on ICDAR 2105 Video dataset.

Methods	Recall	Precision	F-measure
Proposed Text Boxes	85.7	90.2	87.8
Li et al.[38]	65.8	28.6	39.8
Liao et al.[39]	68.8	62.3	65.3
Dey et al.[40]	52.5	42.6	47.0
Yin et al.[41]	55.7	66.3	60.5
Epshtein et al.[42]	60.2	51.1	55.3
Zhao et al.[43]	34.5	22.7	27.4

Table III. Evaluation of the proposed and existing methods on YVT Video dataset.

Methods	Recall	Precision	F-measure
Proposed Text Boxes	88.5	91.3	89.8
Li et al.[38]	57.0	32.0	41.0
Liao et al.[39]	80.7	78.4	79.5
Dey et al.[40]	65.2	57.8	61.2
Yin et al.[41]	57.4	64.4	60.7
Epshtein et al.[42]	76.0	68.0	72.0
Zhao et al.[43]	41.0	34.0	37.0

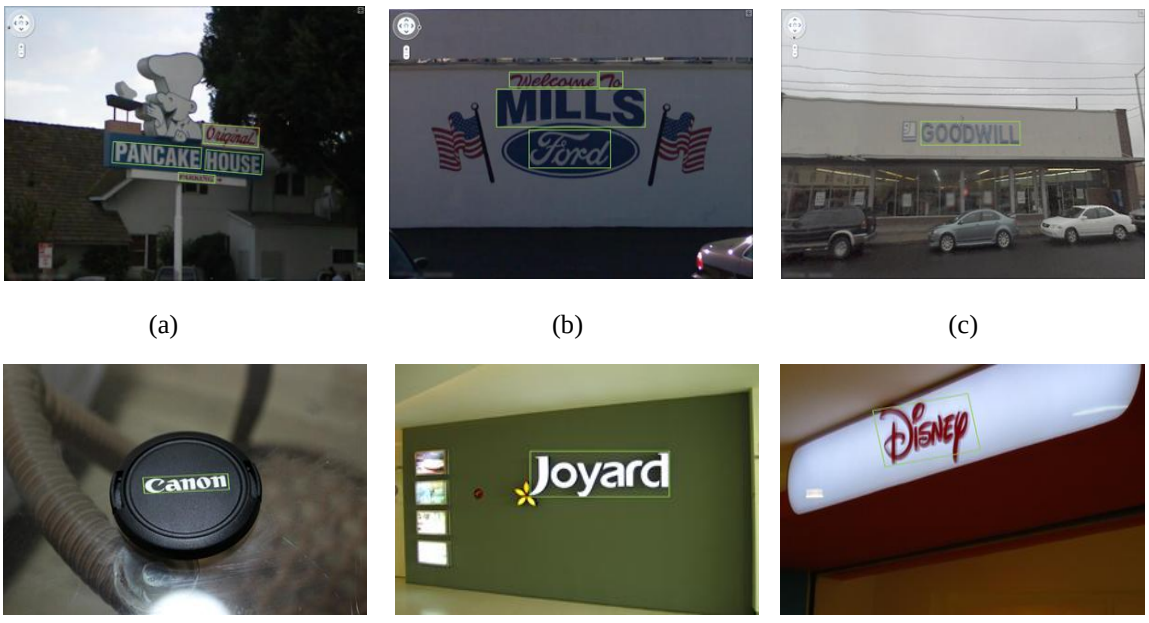
D.2. Experiments on SVT and MSRA Scene Datasets.

Experiments and results conducted by our proposed method on SVT and MSRA Scene datasets are shown in Fig.2 and Table IV and V gives the results in terms of Recall, Precision and F- measure and it is also compared with other existing methods and we can notice the our proposed method illustrations the improved performance than methods in terms Recall, Precision and F-measure. Fig.2. shows the examples of text detection results of proposed method for SVT scene and MSRA scene datasets. For the purpose of comparison with our proposed method we have consider other existing methods such as Li et al. [38], Liao et al. [39], Dey et al. [40], Yin et al. [41], Epshtein et al. [42] and Zhao et al. [43].

(a) (b) (c)

(d) (e) (f)

Fig. 1. Illustrations of text detection results of proposed method on for ICDAR 2105 Videos are (a), (b), (c) and YVT videos datasets are (d), (e), (f).



(a) (b) (c)

(d) (e) (f)

Fig. 2. Illustrations of text detection results of proposed method on for SVT scene are (a), (b), (c) and MSRA scene datasets are (d), (e), (f).

(a) (b) (c)

(d) (e) (f)

Fig. 3. Illustrations of text detection results of proposed method on for our own multilingual datasets are (a), (b), (c), (d), (e), (f).

Table IV. Evaluation of the proposed and existing methods on SVT Scene dataset.

Methods	Recall	Precision	F-measure
Proposed Text Boxes	72.9	78.4	75.5
Li et al.[38]	45.3	26.7	33.6
Liao et al.[39]	67.2	60.8	63.8
Dey et al.[40]	68.0	55.0	61.0
Yin et al.[41]	38.1	50.7	43.5
Epshtein et al.[42]	32.8	48.5	39.1
Zhao et al.[43]	43.6	36.2	39.5

Table V. Evaluation of the proposed and existing methods on MSRA Scene dataset.

Methods	Recall	Precision	F-measure
Proposed Text Boxes	82.2	86.1	84.1
Li et al.[38]	65.0	26.0	37.0
Liao et al.[39]	68.8	65.3	67.0
Dey et al.[40]	85.0	52.0	65.0
Yin et al.[41]	63.0	81.0	71.0
Epshtein et al.[42]	30.6	24.1	26.9
Zhao et al.[43]	69.0	34.0	46.0

D.3. Experiments on our own multilingual Datasets.

Experiments and results conducted by our proposed method on our own Multilingual video datasets are shown in Fig.3 and Table VI gives the results in terms of Recall, Precision and F- measure and it is also compared with other existing methods and we can notice the our proposed method illustrations the improved performance than existing methods in terms Recall, Precision and F-measure. Fig.3. shows the examples of text detection results of proposed method for our own multilingual videos datasets. For the purpose of comparison with our proposed method we have consider other existing methods such as Li et al. [38], Liao et al. [39], Dey et al. [40], Yin et al. [41], Epshtein et al. [42] and Zhao et al. [43].

Table VI. Evaluation of the proposed and existing methods on Own Multilingual dataset.

Methods	Recall	Precision	F-measure
Proposed Text Boxes	87.8	90.5	89.1
Li et al.[38]	43.5	30.4	35.8
Liao et al.[39]	65.4	57.3	61.0

Dey et al.[40]	63.5	55.2	59.0
Yin et al.[41]	50.7	70.5	58.9
Epshtein et al.[42]	49.3	54.2	51.6
Zhao et al.[43]	56.2	39.4	46.3

E. Text Recognition Approach

In the phase of text recognition we consider our proposed method on video that is comprised of our own multilingual datasets of 30 videos and standard benchmark datasets like, ICDAR-2015 of 23 videos, ICDAR 2011 and SVT scene datasets of 150 images each. According to the ground truth and testing data described in the database, we evaluate the proposed method for both character and word levels in all the experiments consider during this work. We have used the cropped character images from ICDAR datasets around 100 million images and synthetic characters and Char74k. From our own multilingual South Indian we consider 600 words from each script and total of 3000 word images. From ICDAR-2015 we have used 400 word images, and from ICDAR 2011 1600 word images and SVT we have used 1500 word images, totally 3500 word images. For the purpose of comparison with our proposed method we have consider other existing methods such as S Roy et al. [44], S J Lee et al. [45] and M Jaderberg et al. [8].

Table VI. Evaluation of Recognition rates proposed and existing methods on various datasets for word and character levels (in %). W and C indicates word and character recognition rates, correspondingly.

Methods	ICDAR 2015 Videos		SVT Scene		ICDAR 2011 Scene	
	C	W	C	W	C	W
S Roy[44]	66.4	53.4	53.8	53.8	59.2	59.2
S J Lee[45]	82.3	76.4	72.4	64.2	75.5	71.3
M Jaderberg [8]	84.8	78.4	78.43	72.3	80.4	74.4
Proposed	89.3	85.6	86.2	82.9	91.5	87.4

Table VII. Evaluation of Recognition rates proposed and existing methods on our own multilingual South Indian datasets for word and character levels (in %). W and C indicates word and character recognition rates, correspondingly.

Methods	Kannada		Telugu		Tamil		Malayalam		All	
	C	W	C	W	C	W	C	W	C	W
S Roy[44]	53.0	48.2	53.4	47.0	54.6	52.7	52.8	51.0	13.7	10.0
S J Lee [45]	50.7	43.2	45.8	41.9	50.2	46.9	53.9	51.5	10.6	8.8
M Jaderberg [8]	56.3	51.8	53.7	51.2	56.3	52.5	56.2	52.4	14.2	10.6
Proposed	67.4	56.3	58.7	54.4	64.5	59.4	65.4	58.6	32.5	25.2

V. CONCLUSION AND FUTURE SCOPE

In this work we have presented text detection method that is built on the deep neural networks which consists of TextBoxes. The proposed method is capable of detecting text in video frames with high accuracy and efficiency when compared state-of-art-methods. In parallel we have also proposed text candidate detection and recognition method for video frames by extracting words based on Automatic Window size prediction for detection of words and to find the direction of words we use principal component analysis and convolution layers of the deep neural network for training character images. Experimentations are done on standard datasets and our own multilingual datasets for both detection and recognition and achieved better results. In future extend our work to improve the performance and to design a multilingual OCR.

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. Pooja and R Dhir, “Video Text Extraction and Recognition: A survey”, 2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), IEEE, Chennai, pp. 1366-1373, 2016.
2. M Lu, H Sun, J Chu,X Huang and J Yu, “A Novel Approach for Video Detection and Recognition based on a Corner response Feature Map and Transferred Deep Convolution Neural Network ”,in IEE Access, vol. 6, pp. 40198-40211, 2018.
3. B Epshtein, E Ofek and Y Wexler, “Detecting Text in Natural Scenes with Stroke Width Transform”, 2010 IEEE Computer Society Confernece on Computer Vision and Pattern Recognition, San Fransico, CA, pp. 2963-2970, 2010.
4. L Neumann and J Matas, “A Method for Text Localization and Recognition inReal-Worlsd Images”, In: Kimmel R., Klette R., Sugimoto A. (eds).Computer Visison –ACCV 2010. Lecture Notes in Computer Science, Springer, Berlin, Heildelberg, vol 6494, 2010.
5. B Shi, B Xiang and J Serge, “Detecting Oriented Text in Natural Images By Linking Segments”, 2017 IEEE Conference on Computer Vision nad Pattern recognition (CVPR), pp. 3482-3490, 2017.
6. Z Xinyu, Y Cong, He Wen,Yuzhi W, Shuchang Z, Weiren He and J Linag, “EAST: An Efficent and Accurate Scene Text Detector”, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, Hi, pp. 2642-2651, 2017.

7. M Jaderberg, A Vedaldi, and A Zisserman, " Deep Features for Text Spotting", In: Fleet D., Pajadla T., Schiele B., Tuytelaras T. (eds) Computer Vision- ECCV 2014. Lecture Notes in Computer Science, Springer, Cham, vol 8692, 2014.
8. M Jaderberg, Simonyan K, A Vedaldi and A Zisserman, " Reading Text in the Wild with Convolution Neural Networks", Internation Journal of Computer Vision Nad Pattern Recognition (CVPR), pp. 1-20, 2016.
9. R Grishick, J Donahue, T Darrell and J Malik, "Rich Feature Hierarchies For Accurate Object Detection and Semantic Segmentation", 2014 IEEE Conference on Computer Vision and Pattern Recgontion, Colombus OH, pp.580-587, 2014.
10. K HE, X Zhang, S Ren and J Sun, "Spatial Pyramid Pooling in Deep Convolution Neural Networks for Visual Recognition", In: Fleet D., Pajadla T., Schiele B., Tuytelaars T. (eds) Computer Vision –ECCV 2014. Lecture Notes in Computer science Springer, Cham, vol. 8691, 2014.
11. Elizaveta K M, Vsevolod I and Trifalenkov., "Text Clustering as Graph Community Detection", 8th Annual International Conference on Biologically Inspired Cognitive Architectures, Procedia Computer Science, Elsevier, vol 123, pp. 271-277, 2018.
12. Shangbang L, JiTextyR, Wenjie Z, Xin He, Wenhao Wu and Cong Yo., "TextSnake: A Felixble Representation for Detecting Text of Arbitary Shapes", Computer Vison and Pattern Recognition (CVPR), ECCV 2018.
13. Y Nagaoka, T Miyazaki, Y Sugaya and S Omachi., "Text Detection by Faster R-CNN with Multiple Region Proposal Networks", 2017 14th International Conference on Document Analysis and Recognition (ICDAR), Kyoto, pp. 15-20, 2017.
14. L M Francis and N Sreenath., "TEDLESS – Text Detection using least square SVM from Natrual Scene Images", Journal of King Saud University – Computer and Information Sceience, Elsevier 2017.
15. P Shivakumara, R Raghavendra, L Qin, K B Raja, Tong Lu and U Pal., "A New Multi-Modal Approach to bib number/text detecion and Recognition in Marathon Images", Pattern Recognition, Elsevier, Vol. 61, pp. 479-491, 2017.
16. V N M Aradhya and M S Pavithre., "A Comprehensive of Transforms, Gabor filter and k-menas clustering for text detectionin images and videos", Appiled Computing and Informatics, Elsevier, Vol 12, issue 2, pp. 109-116, 2016.
17. Y Wang, LPeng and A Wang., "A Multi-Stage Method for Chinese text Detectionin News Videos", 20th Interbnational Conference on Knowledge Based and Intelligent and Engineering Systems, Procedia Computer Science, Elseiver, Vol 96, pp.1409-1417, 2016.
18. M Valdenegro-Toro, P Ploger, S Eickeler and L Konya., "Histograms of Stroke Widths for Multi-Script Text detection and verification in Road Scenes", 9th IFAC Symposium on Intelligent Autonomus Vehicles IAV 2016,Germany, Elseiver, Vol 49,Issue 15, pp.100-107,2016.
19. A B Ayed, M B Halima and A M Alimi., "MapReduce Based Text Detection in Big Data Natural Scene Videos" INNS Conference on Big Data 2015 Program san fransico, CA, USA, Procedia Computer Science, Elsevier, Vol 53, pp. 216-223, 2015.
20. Lei Sun, Qiang Huo, Wei Jia and Kai Chen., "A Robust approach for Text Detection from natural scene images", Pattern Recognition, Elesiver, Vol 48, Issue 9, pp. 2906-2920, 2015.
21. X Wang, Y Song, Y Zhang and J Xin., "Natural Sceen Text Detetcion with Muti-Layer Segemntation and Higer Order Conditional Random Fields based Analysis", Pattern Recogniton Letters, ELseiver, Vol 60-61, pp.41-47,2015.
22. C Shi, C Wang, B Xihao, Y Zhang and S Gao., "Scene Text Detetion using Graph Model built upon Maximally Stable Extermal Regions",Pattern Recogniton Letters, Elsevier, Vol 34, Issue 2, pp. 107-116,2013.
23. V N Manjunath Aradhya, M S Pavithra and C Naveena., "A Robust Multilingual Text Detection Approach Based on Transforms and Wavelet Entropy" Procedia Technology, Elsevier, Vol. 4, pp. 232-237, 2012.
24. P Shivakumara, W Huang, T Q Phan and C L Tan., "Accurate Video Text Detction Through Classification of Low and High Contrast Images", Pattern Recognition, Elsevier, Vol. 43, Issue 6, pp. 2165-2185, 2010.
25. M Anthimopoulos, B Gatos and I Pratikakis., "A Two-Stage Scheme for Text Detection in Video Images", Image and Vision Computing, Elsevier, Vol. 28, Issue 29, pp. 1413-1426, 2010.
26. Qixiang Ye, Q Huang, WGao and D Zhao., "Fast and Roubst Text Detection in Images and Video Frames", Image and Vision Computing, Elsevier, Vol. 23, Issue 6, pp. 565-575, 2005.
27. C W Lee,K Jung and H J Kim.,"Automatic Text Detection and Removal in Video Sequences", Pattern Recognition Letters, Elsevier, Vol.24, Issue 15, pp. 2607-2623, 2003.
28. A K Bhunia, G Kumar, P P Roy, R Balasubramanian and Umapada pal ., "Text Recognition in Scene Image and Video Frame using Color Channel Seletion" , Multimedia Tools and Applications, Springer, Vol. 77, Issue 7, pp. 8551-8578, 2017.
29. S Shetty, A S Devadiga, S S Chakravarthy and K A V Kumar., "Ote-Ocr based Text Recogniton and Extraction from Video Frames", 2014 IEEE 8th International Conference on Intelligent Systems and Control (ISCO), pp. 229-232, 2014.
30. X Rong, C Yi, X Yang and Y Tian., "Scene Text Recognition in Multiple Frames based on Tracking", 2104 IEEE International Conference on Multimedia and Expo (ICME), pp. 1-6, 2014.
31. W Ohyama, S Iwata, T Wakaba yashi and F Kimura., "Detection and Recognition of Arabic Text in Video Frames", In Proceedings of 6th International Workshop on Multilingual OCR, MOCR, Vol. 7, pp. 20-24, 2017.
32. P Shivakumara, Wu Liang, Lu Tong, T C Lim, M Blumenstein and B S Anami., "Fractals based multi-oriented Text Detection system for Recogniton in Mobile Video Images", Pattern Recognition, Vol 68, pp. 158-174., 2017.
33. D Cheng, J M Odobez and H Bourlard., " Text Detection and Recognition in Images and Video Frames", Pattern Recogniton, Vol. 37, Issue 3,pp. 595-608, 2004.
34. M Liao, B Shi, X Bai, X Wang and W Liu., " TextBoxes: A Fast Text Detector with Single deep Neural Network ", Proceedins of the Thirty-First AAAI Conference on Artifical Intelligence (AAAI-17), 2017.
35. W Liu, D Anguelov, D Erhan, C Szegedy, S Reed, C Yang Fu and A C Berg., "SSD: Single Shot MultiBox Detector", Computer Vision and Patter Recognition (CVPR), ECCV 2016.
36. K S Raghunandan, B M Chethan kumar, G Hemantha Kumar, Sunil C., "Convolutional Neural Network based deeo Features for Text Recogniton inMulti-Type Images", 2018 International conference on Advances Computing, Communications and Informatics, Banglore, pp. 502-507, 2018.
37. J Gu, Z Wang, J Kuen, L Ma, A Shahroudy, B Shuai, T Liu, X Wang, Li Wang, G Wang, J Cai and T Chen., "Recent Advances in Convolutional Neural Networks", Computer Vision and Pattern Recognition (CVPR), 2017.
38. H Li, D Doermann and O. Kia., "Automatic Text Detection and Tracking in Digtal Video", IEEE Transactions on Image Processing, pp. 147-156, 2000.
39. M Liao, B Shi, X Bai, X Wang and W Liu., "Text Boxes: A Fast Text Detector with Single Deep Neural Network", In Proceedings of AAAI-17, pp 4161-4167, 2017.
40. S Dey, P Shivakumara, K S Raghunandan, U Pal, T Lu, G H Kumar and C S Chan., "Script Independent approach for Multi-Oriented Text Detection in Scene Images", NEUCO, pp. 96-112, 2017.
41. X C Yin, X Yin, K Huang, and H W Hao., "A Robust Text detection inNatural Scene Images", IEEE Transactions on PAMI, pp. 970-983, 2014.
42. B Epshtein, E Ofek and Y Wexler., " Detecting Text in Natural Scenes with Stroke Width Transform", In Proceedings of Computer Vision and Patteren Recognition (CVPR), pp. 2963-2970, 2010.
43. X Zhao, K H Lin, Y Fu,Y Liu, and T S Huang., "Text From Corners: A novel Approach to Detect Text and caption in Videos", IEEE Transcations on Iamge Processing, pp.790-799, 2011.
44. S Roy, P P Roy, P Shivakumara, G Louloudis and C L Tan., "HMM-based Multi Oriented Text Recognition in Natural Scene Images", 2013 2nd IAPR Asian Confernece on Pattern Recognition, pp. 288-292, 2013.
45. S J Lee and S W Kim., "Recognition of Slab Identification Numbers using Deep Convolution Neural Networks", 2016 15th IEEE International Conference on Machine Leraning and Applications (ICMLA), pp. 718-721, 2016.

AUTHORS PROFILE

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.

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.

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 eight 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. from University 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.