Deep Graph Learning Based Approach for Identification of Text in Scene Video

Mortha Manasa Devi, Dr.M.Seetha, Dr.S.Viswanadharaju
Scholar/JNTUH, Professor/GNITS, Professor/JNTU-Jagtial

Abstract— *Content based analysis, retrieval, searching of scene video has become a key area under computer vision. Apart from indexing and retrieval of videos, demands for video analysis to monitor illegal videos have revolutionized the text detection problem. Because of complex background, low contrast, illuminated, variable font sizes, traditional approach of video based Optical Character Recognition (OCR) system performs satisfactory to detect the text from video. Later, two state-of the-art methods like SIFT and MSER outperformed to detect the text in video but both of these methods fails to detect with complex background. The proposed architecture utilizes the deep graph learning model to detect and identify the scene text from video in two stages. First, regions of similar nature are extracted from the frames by applying undirected graphs. Second, the extracted regions are fed to the learning model to obtain the features which are convolved with internal layers to find the probability of existence of text by calculating the gradients and gray level contrast between text and background. Compared to the conventional detection methods like SIFT and MSER, the detection rate based on deep graph learning can reach 90%. Experimental results show that proposed method is effective compared to two state-of-the-art methods SIFT and MSER.*

Index Terms—***Deep Learning, Graphs, Video, Scene text, CNN, Feature maps.***

I. INTRODUCTION

With the advent of video indexing, searching and automatic annotation, text in video have occupied the recent research efforts. Video text provides the semantic information about the content which promotes content based video analysis. Text plays a major source of information which can be used in diverse applications like video indexing, merchandise movement, de-identification of medical images, geo-coding, assistant to visually impair. Therefore, detection of text from natural scene videos has become popular and essential research in computer vision. Even though many authors have worked on this, however, a succession of challenges may still be encountered because of variable size text in scenes, complex backgrounds, and imperfect image due blur, distortion, and low contrast. All recent methods are built on deep learning models which overcome from designing the traditional way of hard-coded features. Also, researchers are working on unique datasets which are newly published with more challenging features. Therefore, algorithms are proposed by different authors to tackle specific challenges. Herewith, our proposed method also detects the text with more efficient approach by applying the graphs to increase the performance compared to contemporary methods.

II. EQUIVALENT WORK

Researches on detection and identification of scene text in video have been conducted for decades [12]. Recent advancement in pattern recognition and computer vision has changed the evolution of text detection in video. [11] Drastically improved the detection and identification of text by convolution neural network (CNN). However, because of variable font size, style, occlusion, low contrast, blurred, it is difficult to detect all kinds of text. Wang et. al. have utilized sliding window method to sense the text in video and later optical character recognition(OCR) is utilized to identify the text[4]. [11] detected the text with Maximally Stable Extremal Regions (MSER) and fed to deep CNN to recognize. Shi et. al. proposed the combination of CNN and recurrent neural network to train endways system for detection and recognition of text in video[8]. Zhao et. al. used Harris corner points to find the intensity variation in the corner with some heuristic rules to detect the text from scene video[9]. Many authors have proposed independent methods for text detection from scene images and recognition of detected text. In order to utilize the continuous trainable model, it is desirable that system should not only end-to-end but it should also be trainable model from input scene text video to output text. In this case, Lu et. al constructed transferred deep CNN classifiers from VGG16[2], ResNet50[2] with a series of strategies which has achieved tremendous performance[1]. Feng et al. proposed continuous fuzzy systems with variable time delay which provides the method for nonlinear systems [6]. With the advent of deep learning, CNN based text detection in video has been widely used and outperformed compared to SIFT[16] and MSER [7]. However, CNN based approaches are not found to be effective because of their end-to-end system which uses number of convolution layers to predict the corresponding text in video [19]. The proposed approach utilizes the deep graph model which directly works on the input video frames to extract the text regions which is again fed to the deep model to detect the text by finding the probability of occurrence of text in the region. The detected text region is compared with the traditional state-of-the-art methods SIFT and MSER and compared the accuracy with respect to the performance on ICDAR 2015 video text datasets [15]. Our method outperforms with 90% accuracy on
above mentioned datasets. [6] proposed a framework which detects the multi oriented, horizontal and non-horizontal video scene text by applying the skeletonization which has proved to me effective and verified by Hidden Markov Models(HMM).

III. PROPOSED METHOD

The proposed methodology utilizes the deep learning and graph based approach to detect the text in video. The graph based approach is responsible to extract the regions from the frame which is again processed to find the probability of occurrence of text in the region. The novelty of the proposed method is extracting the regions using non-linear data structure consisting of nodes and vertices. First, video is divided into frames, and then each frame is fed to the learning model where model computes the regions by considering pixels as nodes and connection between node to node as edge. The frame is translated to graph, the distance between the pixel elements is calculated and elements having minimum distance are connected to neighboring pixels to form a region. Regions formed with minimum distance nature are extracted and probability is calculated to find whether the region contains text or non-text. The occurrence of text is found by computing gradients of pixels in the regions. If the gradient value is high, then that region is considered to feed to the learning model to learn the features. Figure 1 show the process of region extraction.

Deep Graph Learning

Deep graph learning is a deep neural net which includes an undirected graph as an internal layer to extract the areas of interest from the inout scene video frames. The objective of this learning model is to utilize the deep model without hard-coding the features which was done with traditional approaches. Figure 1 shows the flowchart for identifying the text in video by traversing through the deep layers for regions extraction, feature extraction and classification. We train the model by taking the input frames and fed to the layers which extracts the regions by applying graph approach. Later, regions are fed to the convolution layers to extract the features which are again connected to fully connected layer to classify the text from non-text. Features are extracted by passing the regions through network shown in figure 3.

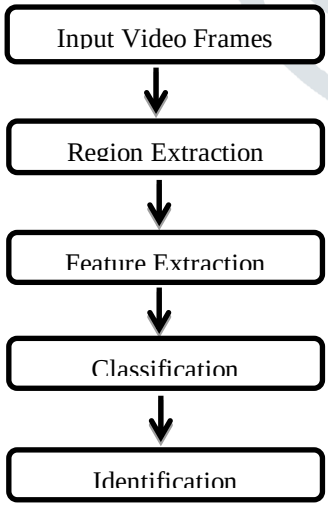


Fig 1. Flowchart for Text Detection in Video

Region Extraction

Regions are extracted from the pre-processed video frames by using undirected graph where nodes are pixels of the frame and edges are the connectivity between the nodes. Based on this, input video frame is converted to matrix which is again translated to grapgh as shown in figure 2. Once undirected graph is generated, then minimum distance is computed to find the nodes having similar distance. Later, regions are extracted with similar nature based on the distance.Graph is a collection of pixels of the input video frames where each pixel is a node and relationship between the neighbouring nodes is edge. Minimum distance is computed between the nighbouring nodes and realtion having minimum distance nodes are grouped in one cluster. Based on this approach, the proposed method is divided in two stages. First, pre-processed video frames are sent to the CNN to generate feature maps. These feature maps are segmented frames of original video frames. Deep graph is applied to these feature maps to separate the regions of interest to different groups. From these groups, wach region is convolved with with the filters which finds the gradients and contrast between text and background to detect the text. Second, these regions are fed to the internal layers to classify the text from non-text by computing the probability of occurrence of text in the region. Region having higher probability is considered to be text otherwise non-text.

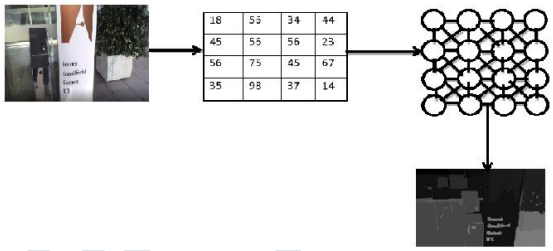


Fig 2. Region Extraction from Input Video Frame

Feature Extraction

Extracted regions from the previous step are fed to the convolution neural network which includes internal layer consists of 32 filters to segment the regions of interest. Then, these regions are pooled to downsample to fixed size. Later, these are connected to fully connected layer which included tangent hyperbolic to find the probability of occurrence of text. The region having high probability is considered as text otherwise non-text. The flow of layers to extract the features are shown in figure 3.

Experimental Results

To compare the results of the proposed methodology, we have used confision matrix to evaluate the performace with state-of –the-art methods SIFT and MSER. We took datasets from ICDAR 2015 video text dataset which contains 25 videos. Table 1 shows the text detection methods and their performace. The proposed method shows thebetter performace than MSER and SIFT. We performed our experiments with python , opencv and tensorflow. Each video from the dataset contains approximately 900 images where randomly chosen for training and teting data. Deep learning model is constructed to train the videos to detect the text by using graph and feature based. Series of layers are used to extact the features, then, the features are applied with tangent hyperbolic function to find the probability of occurrence of text in the region. We have tested our methodology and compared with SIFT and MSER. Recall and Precision is computed to find the overall accuracy of text detection on ICDAR 2015 datasets. Recall is performed by finding percentage of total text regions detected correctly where precision is performed by finding the percentage of total regions detected which are text. Based on these

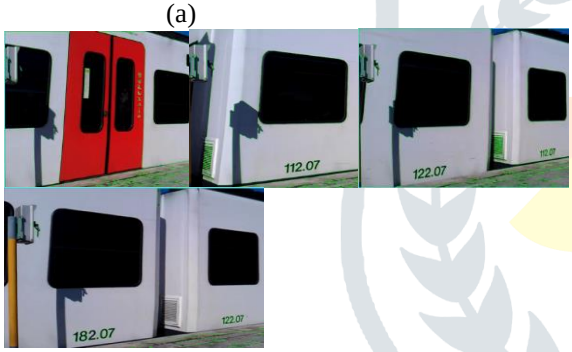
assessment parameters , we have tested on ICDAR 2015 video text datasets by using our method, SIFT and MSER. Table 1 shows the results and comparision of accuracy.

Table 1. Experimental Results of Proposed Method

Method	Recall	Precision	Accuracy
Proposed Method	**0.93**	**0.89**	**0.90**
SIFT	0.86	0.84	0.85
MSER	0.72	0.67	0.65

Compared with traditional designed features in [15] , our constructed deep graph model can learn more descriminative features.

(a)



(b)

(c)

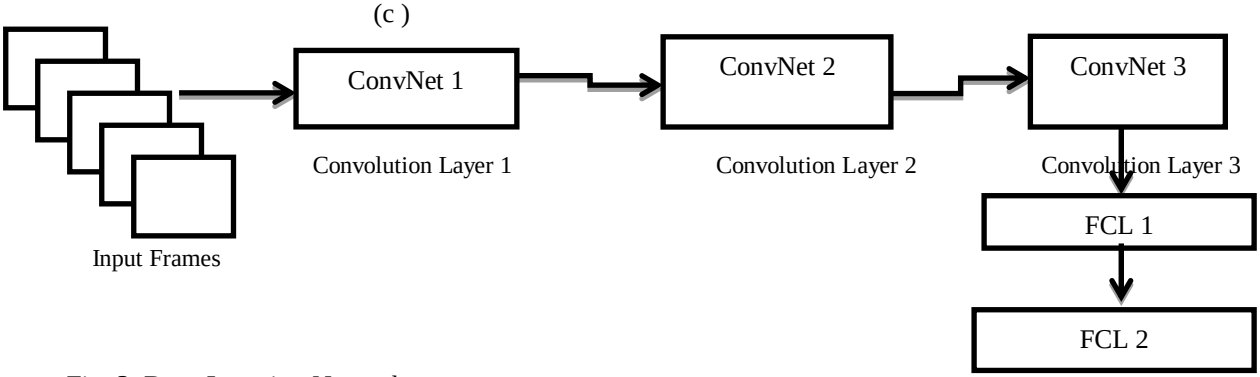


Fig. 3. Deep Learning Network

Fig.4. Text Detection results of the proposed method on ICDAR 2015 Dataset

Figure 4 shows sample results obtained from the robust reading ICDAR 2015 competetion. We can see form the (a) that if text is far away, our method cannot detect. Also, there could be more false positives as shown in (b) and (c).

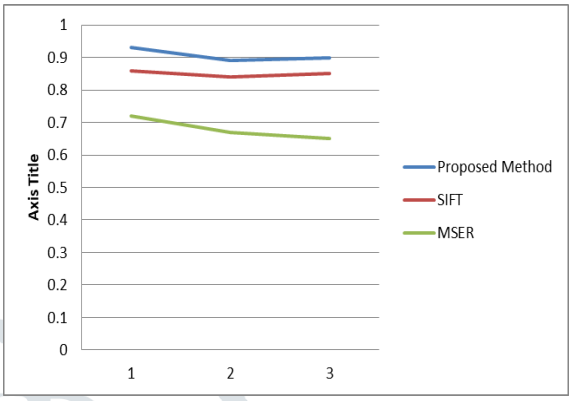


Fig. 5. Comparision of experimental results

IV. CONCLUSION

We proposed the deep graph learning model for text detection in video which utilizes the features maps to extract the regions of interest by using undirected graphs. In order to detect the text, gradients and contrast between text and background is computed to identify the text from extracted regions. Also, deep learning model clarified that te proposed model is more effective than SIFT and MSER feature based dettection methods. The proposed model works in end-to-end system with simple architecture to classify text from non-text. Therefore, detection performace of current method outperforms compared to tradiotnal state-of-the-art methods. However, deep graph learning cannot detect far scaled and blurred text. Hence, we will improve the detection algorithm with most effective learning model for all kinds of scene text in video in the future.

V. REFERECES

[1] Wei Lu, Hongbo Sun, Jinghui CHhu, Xiangdong Huang, Jiexiao Yu, "A Novel Approach for Video Text Detection and Recognition Based on a Corner Response Feature Map and Transferred Deep Convolutional Neural Network",IEEE, pp. 40198-40211,2018.

[2] Lan Wang, Yang, Susu Shan, Feng Su, "Scene Text Detection and Tracking in Video with Background Cues", pp.160-168, International Conference on Multimedia Retrieval (ICMR), ACM, 2018

[3] Chun Yang, Xu-Cheng Yin, Senior Member, IEEE, Wei-Yi Pei, Shu Tian, Ze-Yu Zuo, Chao Zhu, and Junchi Yan," Tracking Based Multi-Orientation Scene Text Detection: A Unified Framework With Dynamic Programming", IEEE Transactions on Image Processing, vol. 26, No.7, pp. 3235-3248, July 2017.

[4] J. Wang and H. Wang, ''A study of 3D model similarity basedonsurfacebipartitegraphmatching,''Eng.Comput.,vol.34,no.1,pp.174–188,2017.

[5] Qixiang Ye and D. Doermann, "Text detection and recognition in imagery: A survey," IEEE Trans. Pattern Analysis Machine Intelligence, vol. 37, no. 7, pp. 1480–1500, Jul. 2015.

[6] Z. Feng and W. X. Zheng, ''Improved stability condition for Takagi–Sugeno fuzzy systems with time-varying delay,'' IEEE Trans. Cybern., vol. 47, no. 3, pp. 661–670, Mar. 2017

[7] Xu-Cheng Yin, Ze-Yu Zuo, Shu Tian, and Cheng-Lin Liu, "Text Detection, Tracking and Recognition in Video: A Comprehensive Survey," IEEE Transactions on Image Processing, vol.25, no. 6, June.2016.

[8] J. Shi, X. Luo, and J. Zhang, ''Dct feature based text capturing and tracking,'' in Proc. Chin. Conf. Pattern Recognit., Nov. 2009, pp. 1–4.

[9] X. Zhao, K.-H. Lin, Y. Fu, Y. Hu, Y. Liu, and T. S. Huang, ''Text from corners: A novel approach to detect text and caption in videos,'' IEEE Image Process., vol. 20, no. 3, pp. 790–799, Mar. 2011.

[10] Yuqi Zhang, Wei Wang, Liang Wang, and Liuan Wang, "Scene Text Recognition with Deeper Convolution Neural Networks," IEEE, ICIP-2015.

[11] Weilin Huang, Yu Q, and Xiaoou Tang, "Robust Scene Text Detection with Convolution Neural Network Induced MSER Trees," Springer, ECCV-2014, pp. 497-511.

[13] Jing Zhang, and RangacharKasturi, "A Novel Text Detection System Based on Character and Link Energies," IEEE Transactions on Image Processing.Vol.23, no. 9, Sep-2014.

[14]Honggang Zhang, Kaili Zhao, yi-Zhe Song, Jun Guo,"Text extraction from natural scene image: A survey", Neurocomputing, Elsevier, pp. 310-323, 2013.

[15] Manasa Mortha, M.Seetha, S.Viswanadha Raju," Detection and Tracking of Text from Video using SIFT and MSER", ICDECT, March, 2019.

[16]ViniVidyadharan, and SubuSurendran, "Automatic Image Registration using SIFT-NCC", Special Issue of International Journal of Computer Applications (0975 – 8887), pp.29-32, June 2012.

[17] Weilin Huang, Yu Q, and Xiaoou Tang, "Robust Scene Text Detection with Convolution Neural Network Induced MSER Trees," Springer, ECCV-2014, pp. 497-511.

[18] Mortha Manasa Devi, M.Seetha, S.Viswanadha Raju," Text Spotting in Video:Recent Progress and FuturTrends", IJRTE, Vol. 7, pp. 116-124, April, 2019.

[19] Tao Wang, David J. Wu, Adam Coates, and Andrew Y. Ng, "End-to-End Text Recognition with Convolution Neural Networks,"ICPR-2012.