IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-5, May 2019 E-ISSN: 2347-2693

Video Text Detection and Recognition Based on a Transferred Deep Convolutional Neural Network

Namrata Choudhary[1*], Kirti Jain[2]

[1,2]Dept.of Information Technology, Lakshmi Narain College of Technology, Bhopal, MP. India

DOI: https://doi.org/10.26438/ijcse/v7i5.718723 | Available online at: www.ijcseonline.org

Accepted: 19/May/2019, Published: 31/May/2019

Abstract- The text presented in videos contains important information for content analysis, indexing, and retrieval of videos. The key technique for extracting this information is to find, verify, and recognize video text in various languages and fonts against complex backgrounds. In this paper, we propose a novel method that transferred deep convolutional neural networks for detecting and recognizing video text. We partition the candidate text regions into candidate text lines by projection analysis using two alternative methods. We develop a novel fuzzy c-means clustering-based separation algorithm to obtain a clean text layer from complex backgrounds so that the text is correctly recognized by commercial optical character recognition software. The proposed method is robust and has good performance on video text detection and recognition, which was evaluated on three publicly available test data sets and on the high-resolution test data set we constructed.

Keywords— Video Text Detection, Recognition, Transferred convolutional neural network, Fuzzy c-means Clustering.

I. INTRODUCTION

With the rapid development of the internet, communication technology, and smart phones, video has become the most popular medium. The number of online videos has increased dramatically because of the convenience of uploading and downloading videos. Accordingly, there is a high demand for efficient indexing, retrieval, and localization of desired content from massive videos. A lot of algorithms have been developed for this purpose [1]. For video indexing and retrieval, the video text can depict the video more directly and accurately compared with low-level perceptual content (such as edge, shape, and texture) and other semantic content (such as face, vehicle, and human action) [2]. Furthermore, the content analysis of video text can be used to monitor illegal videos. On the other hand, after text areas are localized, neighbor-pixel interpolation algorithms can be used to restore images that are blocked by text.

Thus, video text detection and recognition are significant and challenging tasks because of variations in languages, fonts, and complex backgrounds [3]. Generally speaking, video text can be classified into scene text and artificial text [4]. Of these, the latter usually concisely depicts important video content. For instance, captions in news videos usually describe event information, and subtitles in speech videos usually provide core ideas. Thus, in this paper, we focus on the detection and recognition of artificial text in video frames. Video OCR technology [5] generally has similar processing steps, including text detection, localization, extraction, and recognition. The detection step aims to find text regions; the localization step concentrates on the accurate position of text lines.

First, it is difficult to identify text regions accurately and completely because of various languages, fonts, resolutions, and particularly complex backgrounds. For example, edge-based approaches may produce many false positives when the complex background also has a high density of edges. Second, the heuristic constraints and machine learning methods proposed to eliminate false positives for video text are always optimized for specific situations, which reduce the generalizability of these methods. In fact, no matter what the language and font the text has, the component characters are always formed by crosses of strokes in limited space. Therefore, many corners exist [6]. CNNs can learn discriminative features for precise classifications directly from a large amount of diverse raw data. Transfer learning can transfer the knowledge from one specific task to relevant tasks with good performance.

II. TRADITIONAL METHODS FOR TEXT DETECTION

A large number of techniques have been developed by various researches for text detection and recognition in natural scene images. All these techniques are roughly classified into three basic techniques in figure 1:

A. Texture Based Method

This method uses the texture based properties such as Fourier transform, local intensity, filter response and wavelet coefficients for distinguishing the text part and non-text part from the natural images. Region Based.

B. Method:

This method uses the properties like color, intensity and edge similarity for distinguish the text and non-text part in natural images. It is categorized into three types:

1). Edge Based: this method used the edge detector operator to detect the edges of the images. Usually, two types of edge detection methods are applied such as canny and Soble edge operator.

2) Connected Component Based: In this method, character components are identified using clustering and edge detection methods. Maximum Stable Extremal Region is one of the major techniques of this method.

3) Stroke Width: In this method, text features can be identified through stroke of the components. Character components having constant strokes are treated as text and remaining are treated as non-text. Stroke width transform operator is used for this operation in text detection.

C. Hybrid Method

To overcome the limitations of all the above mentioned techniques, the combination of two or more techniques is used known as hybrid technique.

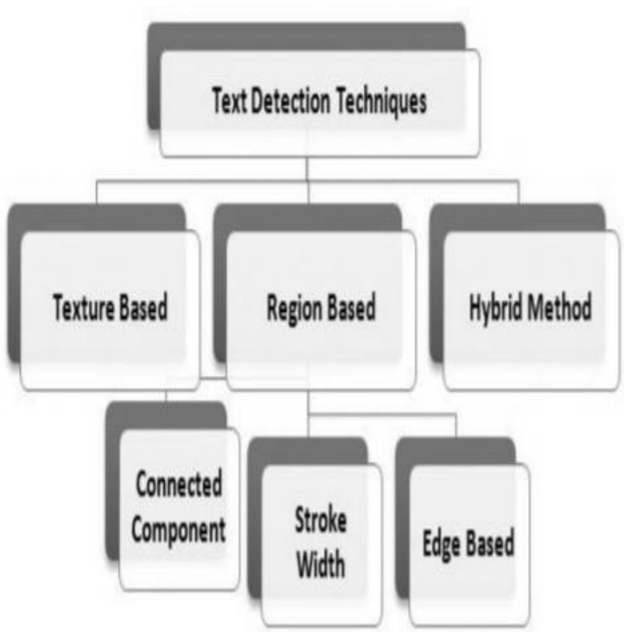

Figure 1: Various types of Text Detection Techniques

Table 1: Describes the Advantages and Disadvantages of Traditional Text Detection Methods

Text Detection Methods	Advantages	Disadvantages
Texture based Method	Can easily handle the noisy image	High computing complexity. Fails in multi orientation text
Region based method • Component Connected method • Stroke Based • Maximum Stable Extremal Region	• Easily handle natural scene images • Low computational cost • Handle the multi-oriented text	• Fails to detect on light variations text images • Very sensitive to noisy images
Hybrid Method	surmount the drawbacks of texture based method and region based method	Complex in implementation

III. DEEP LEARNING IN TEXT DETECTION

In the recent years, due to the increased popularity of deep learning techniques, the use of this technique has become trendy for the task of computer vision and pattern recognition problem in natural scene images. Now a days, many researchers are adopting the deep learning technique for text detection in natural scene images such as conventional neural network, recurrent neural network, fully conventional network, feed forward back propagation neural network. The basic functioning of the convolutional neural network is explained in figure 2.

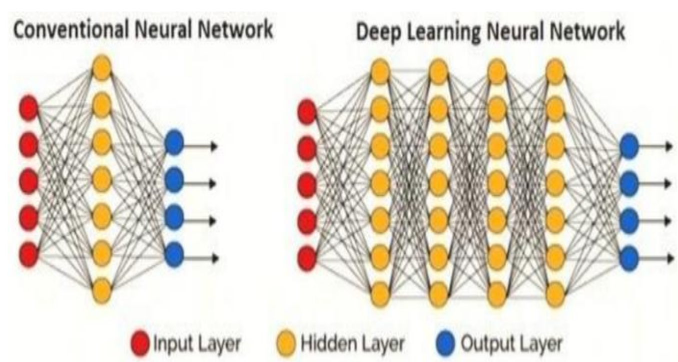

Figure 2: Depicts the architecture of conventional neural network and deep learning neural network

Before 2016, in scene text detection method, the character candidates were identified using hand-craft features. The Convolutional Neural Network/Random forest was used to discriminate the Text / non-text part and to eliminate the false positive in natural images. After 2016, segmentation based method; proposal-based method and hybrid method are used for text detection in natural scene images.

Deep neural network is a term which is widely used these days, consists of multiple hidden layers as compared to simple neural network as shown in figure 3. So, this is the basic reason that the deep learning neural network provides more accurate results as compared to that of simple neural network.

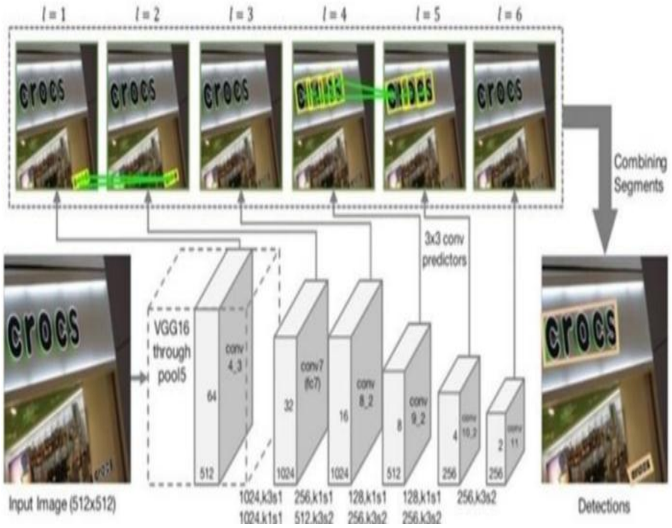

Figure 3: Explains the Deep Neural Network Architecture

IV. PIPELINE PROCESS FOR TEXT EXTRACTION

The overall system for text extraction involves the various steps and the accurateness of every step is mandatory for the accurate text detection. In the process of text detection, the location of detected text is identified using various techniques. Then the identified text needs to get separated from non-text region of the natural scene image. Mainly the pipeline process for text extraction consists of following steps:-

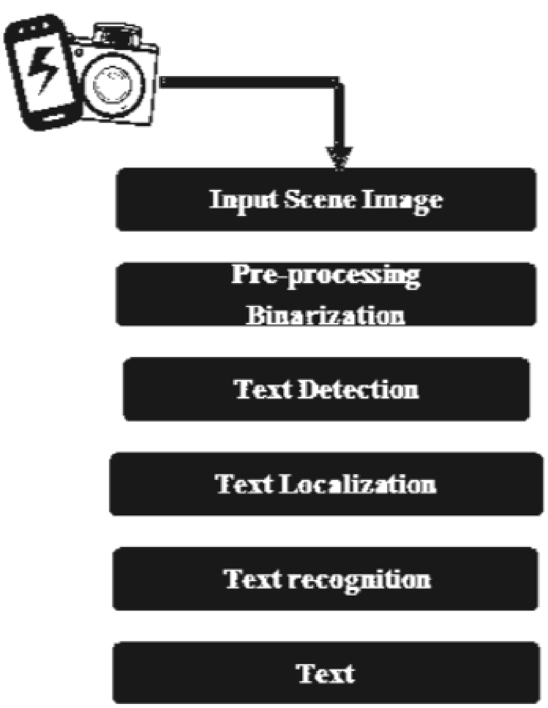

Figure 4: Describes the various steps involved in text extraction

The accuracy of any text extraction technique depends upon the accuracy of text detection technique. So, the text detection in natural images has become significant area of research these days. Here we describe the various steps of text detection applied on real time image.

Figure 5: Depicts the pipeline process for text detection in natural scene images.

V. PROPOSED METHODOLOGY

The proposed approach is composed of four steps: video decoding, text detection, candidate text line localization, and false text line elimination using a deep learning method. First, we utilize the OpenCV library to decode video into frames. Next, we use a corner response feature map detector to obtain candidate text regions. Because there may be multiple text lines in the candidate text region, we then further partition the candidate text lines using two alternative methods. For the first method, candidate text lines are partitioned through projection analysis onto the contours of candidate text regions. If the first method fails, we use a more complicated method, which employs an FCM-based separation method to extract the candidate text layer, converts it to the gray-scale image, and conducts the projection analysis to partition the candidate text lines. In the last step, false text lines are removed by our constructed transferred deep CNN classifiers. The true text lines then undergo FCM-based separation, Otsu binarization, and morphological restoration to obtain OCR-ready binary text. Figure 4 shows the flowchart for the proposed method.

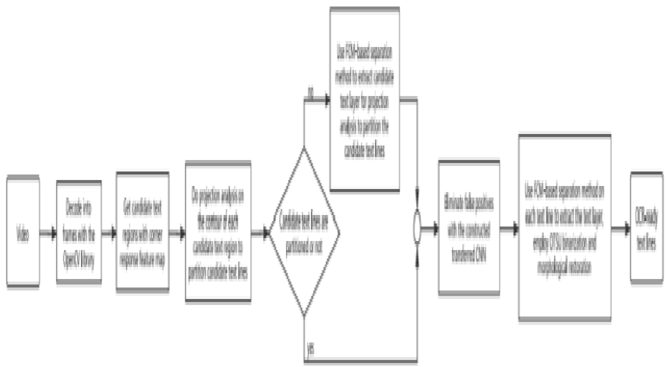

Figure 6: Flowchart for the proposed method

Corner Response Feature Map

Text in videos always provides supplemental information with good readability (especially the captions). The crosses of strokes in characters cause the generation of many corners.

Video text always has a regular distribution of corner points, which the background generally does not have. Compared with other features, such as edge feature, corners are more stable and robust. The detailed mathematical derivation about corners was presented in. Given a gray-scale image *I*, we take an image patch over the window *W*(*x*; *y*), shift it by (u,v), and calculate the change produced by the shift as follows:

$$E(u, v) = \sum_{W} [I(x+u, y+v) - I(x, y)]^2. \tag{1}$$

The first-order Taylor expansion after omitting the Peano remainder term is used to approximate the shifted image as follows:

$$I(x+u, y+v) \approx I(x, y) + \begin{bmatrix} I_x(x, y) & I_y(x, y) \end{bmatrix} \begin{bmatrix} u & v \end{bmatrix}^T \tag{2}$$

Where *Ix* and *Iy* denote first-order partial derivatives in *x* and *y* directions, respectively, substituting approximation (2) into (1) yields:

$$E(u, v) = \begin{bmatrix} u & v \end{bmatrix} M \begin{bmatrix} u \\ v \end{bmatrix} \tag{3}$$

Where *M* is the following Hessian matrix:

$$M = \begin{bmatrix} \sum_W (I_x(x, y))^2 & \sum_W I_x(x, y) I_y(x, y) \\ \sum_W I_x(x, y) I_y(x, y) & \sum_W (I_y(x, y))^2 \end{bmatrix} \tag{4}$$

If the two eigenvalues of *M* are large and distinct positive values, a shift in any direction will cause a significant increase, and a corner can be determined.

Figure 7: Sample video frames and corresponding CRMs after gray-scale morphological processing

Text Detection

OpenCV is an open-source computer vision library that can be used to process videos and images. First, we use it to decode the video into frames through the CVQuery Frame function. Considering human visual characteristics, the video texts always last at least 2 seconds. Therefore, we grab one frame per second for video text detection so that no video text is missed.

We obtain the CRM of original frame according to (4). Regions of higher brightness always correspond to video texts. We then apply a series of gray-scale morphological operations to enhance the text regions and suppress the non-text regions.

A close operation is utilized to remove the dark points that belong to the background in CRM, and then the tophat operation is used to enhance the bright text areas. The combination of the two operations makes the text regions more distinct and complete. The CRMs after gray-scale morphological processing are presented in figure 7.

After the gray-scale morphological processing, Gaussian filtering is used to smooth the CRM, which contributes to the completeness of text regions to be detected. In order to form reasonable candidate text regions, we propose a binarization method with an adaptive threshold.

VI. SIMULATION RESULT

The performance of the proposed method is evaluated using three publicly available test datasets and our proposed test dataset. The three public datasets are the Microsoft common test set, TV news test set, and YouTube test set.

The first dataset contains 45 pictures of low resolution and poor quality, which is not up-to-date. The other two datasets contain high-resolution pictures. However, the size of the two datasets is too small to support further research. Our constructed dataset consists of more than 6,000 typical video frames of high resolution and high quality, about 25,000 text lines, and 42,000 negative samples. These frames are collected from various sources, including movies, cartoons, and TV shows. We sampled 2,000 video frames randomly and used them as the proposed test dataset.

We performed our experiments using Python with the Theano backend and CCC with the OpenCV library. The hardware configuration includes an NVIDIA Geforce GTX 1080Ti with 11-GB GPU memory, an AMD Ryzen5 1400@3.20GHz_4 processor with 64-GB RAM. We resized the candidate text line images into the following input sizes: 224×224 for TVGG and TRESNET, 299×299 for TINCEPTION. In our constructed dataset, 2,000 images are randomly chosen as the test data. For the rest of the images, 80% are randomly selected for training, and the remaining 20% are selected for validation. We adopted the pixel-based evaluation method in, and the experimental results are shown in Table 2. The results show that our methods achieve good performance on a wide range of videos, and our TVGG based method performs best. Therefore, we chose the TVGG based method to compare with several state-of-the-art methods on three public test sets.

Table 2: Experimental results on the proposed test set

Method	Recall	Precision	FI-measure
Our TVGG based Method	0.88	0.83	0.85
Our TRENET based Method	0.88	0.82	0.85
Our TINCEPTION Based Method	0.87	0.82	0.84

VII. CONCLUSION

In this paper, we focus on two different techniques of text detection using traditional methods and the deep learning methods. The traditional methods use the texture based, connected component based and hybrid approaches to detect the text in natural scene images. All these techniques have their own advantages and disadvantages. The study reveals that most of the researchers have applied the deep learning methods and traditional methods for text detection mainly on the horizontal text and near horizontal text, a very few on the curved text. It is concluded that a significant amount of work needs to be done for multi-oriented text detection in natural scene images. The author believes that the problem of text detection in natural scene images is incomplete without text detection in horizontal, non-horizontal and multi-oriented text in Video.

REFRENCES

[1] Y. A. Aslandogan and C. T. Yu, ‘‘Techniques and systems for image and video retrieval,’’ IEEE Trans. Knowl. Data Eng., vol. 11, no. 1, pp. 56–63, Jan. 1999.

[2] H. Bhaskar and L. Mihaylova, ‘‘Combined feature-level video indexing using block-based motion estimation,’’ in Proc. Conf. Inf. Fusion, Jul. 2010, pp. 1–8.

[3] W. Hu, N. Xie, L. Li, X. Zeng, and S. Maybank, ‘‘A survey on visual content-based video indexing and retrieval,’’ IEEE Trans. Syst., Man, Cybern. C, Appl. Rev., vol. 41, no. 6, pp. 797–819, Nov. 2011.

[4] K. Jung, K. I. Kim, and A. K. Jain, ‘‘Text information extraction in images and video: A survey,’’ Pattern Recognit., vol. 37, no. 5, pp. 977–997, 2004.

[5] M. Khodadadi and A. Behrad, ‘‘Text localization, extraction and inpainting in color images,’’ in Proc. Iranian Conf. Elect. Eng., May 2012, pp. 1035–1040.

[6] A. Mosleh, N. Bouguila, and A. B. Hamza, ‘‘Automatic inpainting scheme for video text detection and removal,’’ IEEE Trans. Image Process., vol. 22, no. 11, pp. 4460–4472, Nov. 2013.

[7] M. Cai, J. Song, and M. R. Lyu, ‘‘A new approach for video text detection,’’ in Proc. Int. Conf. Image Process., vol. 1, Sep. 2002, pp. I-117–I-120.

[8] T. Yusufu, Y. Wang, and X. Fang, ‘‘A video text detection and tracking system,’’ in Proc. IEEE Int. Symp. Multimedia, Dec. 2013, pp. 522–529.

[9] X. Huang, ‘‘A novel video text extraction approach based on Log–Gabor filters,’’ in Proc. Int. Congr. Image Signal Process., vol. 1, Oct. 2011, pp. 474–478.

[10] P. Shivakumara, W. Huang, and C. L. Tan, ‘‘Efficient video text detection using edge features,’’ in Proc. Int. Conf. Pattern Recognit., Dec. 2008, pp. 1–4.

[11] 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.

Author's profile

Ms. Namrata Choudhary pursed Bachelors of Engineering form Lakshmi Narain College of Technology, Bhopal, M.P., India in 2017. She is currently pursuing M.Tech. from Lakshmi Narain College of Technology, Bhopal, M.P., India. She has presented paper in National Conference in 2015 & 2016 held in College.

Mrs. Kirti Jain pursed Bachelors of Engineering and M.Tech. She is currently pursuing Ph.D. and working as Associate Professor in Lakshmi Narain College of Technology. She has been published paper in International and National journals and conferences.

