Computer Vision Approach for Detection and Extraction of Text from Video Frames using Color Continuity, Color Variations and HV Projections Methods

Ramgoapl Segu

Electronics and Communication Dept.
Dayanand Sagar College of Engg.,
Bengaluru, India.
ram.segu@gmail.com

K.Suresh

SEA Engineering College
Bengaluru, India.
ksece1@gmail.com

***Abstract - M*ultimedia data retrieval and indexing is an important aspect for information processing and information extraction. In the field of multimedia data indexing, video data indexing is considered as a challenging task for various type of real-time applications. In order to improve the retrieval and indexing performance, text detection and extraction has gained attraction from research community. In this work, we presented text detection and extraction process from video scene frames. In order to accomplish the desired objective, the complete process is divided into two stages as (a) text detection and (b) text recognition. For text detection, a combined approach is presented where color continuity, color variations are computed and later text localization is performed using horizontal and vertical projections. In the next phase, text recognition is performed where DWT based features are extracted and trained using RBF classifier. The complete experimental study is carried out using MATLAB simulation and tested for open source datasets. The comparative study shows that the proposed approach achieves better performance when compared with the existing techniques.**

Keywords - Multimedia Text Recognition, SIFT, DWT, SVM, RBF

I. INTRODUCTION

Demand of multimedia communication systems is increasing substantially in diverse applications in real-time scenarios. These type of multimedia communication systems includes video and image data for multimedia based applications such as video conferencing, video-on-demand services and infotainment systems. In this field, data is increasing rapidly where data storage and retrieval becomes a tedious task. Conventional techniques for video retrieval systems uses manual annotation system with a small keyword for human reviewer.

However, these techniques are time consuming and provide less accuracy of data indexing. Hence, there is a need to develop an automated system which can be used for the video indexing using automated text detection localization and extraction scheme [1-2].

Automated text detection is widely adopted in various applications such as OCR where document text localization can be performed for page segmentation and document retrieving, video or image database indexing for retrieval applications [3]. Figure 1 shows some samples frames of video text in natural scene images.

(a) Natural scene image (scene text)

(b) Video text image (caption text)

Figure1. Video text images (a) scene text image (b) caption text image

Moreover, extraction of text from natural scene images is also considered as an important task which can be used for vehicle license plate recognition, text-based landmark identification and object identification etc. The text detection process is useful, where text provide unambiguous information about the video content and provides the meaningful keywords about the video content. Generally, two type of texts are found in the video which are called as scene text and caption text and as depicted in figure 1(a) and (b) respectively [4]. Generally, caption texts are superimposed artificially during video editing process which helps to summarize the video content, whereas, scene texts occur naturally during the video capturing. In scene texts, sign, banners etc. can be used to describe the video content. Video caption technique is considered as a promising technique for video data indexing because it provides the rich information about the video as available in the sequence. However, text extraction from video sequences is considered as a challenging task when compared with the document text extraction due to several challenges such as: video sequences suffer from the background complexity which raise the challenge of identification of text region, scene text or captions in the text tend to have lower resolution because these caption texts very smaller to avoid the occlusion in video frames hence these captions cannot be used directly for OCR (Optical character recognition) systems. Moreover, lossy compression-based MPEG method further degrades the quality of video frame [5].

Generally, text detection methods are divided into three main approaches that are called as connected component-based technique [7], edge-based technique [8], and texture-based technique [6]. According to the connected component technique, color quantization and region expansion methods are applied to accumulate the adjacent pixels of similar colors and constructs a matrix of connected components. The connected components don't maintain the information related to the shape and characters due to low contrast and color bleeding issues. Hence, these techniques are not considered as an efficient solution for the text detection. In order to overcome the issue of low –contrast and color bleeding, edge-based techniques are introduced recently. According to these methods, edge map is estimated for the given input data. Later, horizontal and vertical profiles are evaluated for the computed edge map. Although, these methods provide inaccurate performance for complex background data. In order to mitigate this issue, texture based methods are applied where text region is computed as unique contour for the given video frame. In these techniques, fast Fourier transform, wavelet decomposition, discrete cosine transforms and Gabor filter etc. techniques are applied for extracting the frequency domain features. Later, these schemes require classification techniques such as neural network, and SVM (Support Vector Machine) classifier etc. hence, database training is required for this type of techniques. For significant performance, these techniques require huge database for training which includes text and non-text sample images [9].

On the other hand, text extraction and recognition are also considered as an important aspect for the real-time video text detection and extraction. Moreover, applications are suitable for the visually impaired person. Several techniques are presented recently for character recognition from natural video scenes using computer vision application. Conventional OCR systems were design for the scanned documents where text is in the well-formatted and captured under the controlled environment such as typewritten scanned documents etc. hence, character recognition for the natural scene or video sequence is a challenging task due to the dissimilar text size, color variations, font variations, background and lighting conditions. Text extraction from natural scene images and videos is a progressively growing research area, where numerous techniques have been introduced for text recognition such as binarization [10], edge detection [11] and morphological operations [12] etc. Moreover, machine learning based classification approaches such as neural network [13], fuzzy clustering [14] and Bayesian classifier [15] etc... are also widely adopted in this field. These studies show that text extraction from natural scene images is a complex task due to the background complexities, buildings, trees, window frames etc., the complexity issues increases when text extraction processes are applied for the natural scene videos where video content is dynamic. Less amount of work has been carried out for video text detection. In this work, we focus on the natural scene video analysis and develop a novel approach for the text detection from the natural scene videos.

A. Contribution of the Work

In this work, our main aim is to develop a robust scheme for text detection, localization and extraction from natural scene images and videos. In order to achieve the desired outcome, we present a computer vision based scheme where main contribution of the proposed approach are as follows: first of all, we develop a text detection approach in natural scene images with the help of SIFT and curvature based features. In the next phase of the work, we focus on color distribution modeling which is used for extraction of the text region, later, color variation model is applied, followed by a text localization approach, finally, text verification and recognition technique is performed to achieve the desired outcome.

B. Article Organization

The rest of the manuscript contains four section as section II, section III, section IV and section V where we present brief literature in section II, problem formulation and identified solution in section III, experimental study in section IV and section V provides concluding remarks.

II. LITERATURE SURVEY

In this section, we present a brief discussion about the recent techniques in the field of natural scene text and caption text detection using image processing approach. The complete section is divided into two main sub-sections as text detection where we discuss about the various approaches for natural scene image text detection and text recognition where natural scene image text recognition techniques are discussed.

A. Text Detection and Localization

Text detection techniques are based on the edge, texture, motion, color and other features (i.e. stoke width features) to distinguish the background and text and background. Huang et al. [16] used feature extraction process for text detection and localization in natural video scenes. According to this approach, Log-Gabor filters are used for generating the stroke map which helps to reduce the background content. Later, texture feature extraction is applied on each stroke map for locating the text lines and finally, Harris corner detection is applied followed by morphological operations which are used for connecting the corners in the text region and remaining non-text regions are removed. However, this results in false positives due to inappropriate performance for the arbitrary text in the scene. In order to deal with this type of issues, Shivakumara et al. [17] developed an approach for the video text detection. Authors suggested that conventional approaches for text detection are suitable for the horizontal texts but fails to provide the desired performance for arbitrary orientations. To achieve this task, Fourier-Laplacian filtering scheme is applied and then k-means clustering is applied for candidate text region identification and later connected components are extracted for distinguishing between text strings from the neighboring components. Finally, edge density and text straightness are evaluated for eliminating the false positive.

Similar to the work presented in [16], authors Li et al. [18] presented stroke width-based text detection approach. In their work, a unique contrast-enhanced Maximally Stable Extremal Region (MSER) approach is implemented for extracting the character candidates for text region identification and geometric parameter constraints are applied for removing the non-text regions and stroke width generation is applied for removing the false positives. Finally, the obtained MSERs are clustered together and text region is obtained. The MSER shows a promising performance for text region identification. Hence, Shi et al. [19] formulated the text identification problem as bi-label where text and non-text regions are present in the natural text scene. By taking the advantage of MSERs, a graph modelling is developed for combining the multiple information in a single framework. During this process, MSERs are constructed to identify the text and non-text regions. Robust features are extracted based on the color and geometric features which are used for clustering the text region and discard the discontinuities. Final, labels are obtained using graph cut algorithm where a cost function is developed to select the optimal label.

Yang et al. [20] developed a combined scheme for text localization and recognition framework. In this work, text-localization scheme is developed initially where multi-scale text detector is developed for identifying the potential text in the video frame as a fast-localization verification algorithm. in the next phase, the detected text candidates are filtered by using image entropy based filtering scheme and skeleton-based binarization method is also developed for the background and text separation. Finally, false positives are eliminated using stroke-width transform and SVM classifier. For arbitrary orientation, Liang et al. [21] introduced a new approach where arbitrary text orientation and multi-scripts are considered for analysis. Similar to the work presented in [17], in this model, Laplacian convolution and wavelet frequency domain based features low-resolution text pixel enhancement. In this work, stroke width transform is applied for candidate text region detection. Candidate text regions and nearest neighbor clusters are extracted and used in text alignment. Text detection and recognition in natural video scenes is considered as a tedious task where blur, distortion and camera noise can cause performance related issues. Based on these assumptions, Khare et al. [22] presented a novel scheme for identifying the degree of blur in the video/image later, blind deconvolution approach is implemented for improving the edge intensity along with suppression of blurred pixels. Mittal et al. [23] also aimed on the issues of text recondition such as low contrast, and background noise which can degrade the overall performance of the system. In this work, authors presented a new approach for the detection of multi-oriented text in the video frames. In this process, image enhancement is applied initially which uses sub-pixel based mapping for image enhancement. In the next phase, connected component extraction is applied and Histogram of Oriented Moment feature extraction technique is implemented. Next, SVM classifier is used for identification of text or non-text regions and RCNN (Recurrent Neural Network) is applied for text recognition.

Recently, Bhunia et al. [24] introduced color channel based approach for text detection in natural scene images. However, conventional techniques are based on the image binarization which are not suitable for video scene text detection. Hence, authors presented a color channel based scheme for the text recognition. According to this process, a color channel is automatically selected which is further used for text recognition. In order to perform the text recognition, Histogram of Oriented Gradient (HoG) features are extracted which are later processed through the Hidden Markov Model (HMM). Later, multi-class SVM is applied for recognition. This approach significantly reduces computational complexity due to color channel selection.

B. Text Recognition

We explore maxima stable extreme regions along with stroke width transform for detecting candidate text regions. Text alignment is done based on the distance between the nearest neighbor clusters of candidate text regions. In addition, the approach presents a new symmetry driven nearest neighbor for restoring full text lines.

Multi-oriented detection of text without any restrictions on background, alignment, and contrast, with high precision, and recall still remains a difficult task. Most existing methods [5, 9, 19] depending highly on the horizontal text-orientation fail in cases of multi-oriented text fields. Research on curved-text-line detection is even rarer because not only edge-focused methods fail but also due to most of the above mentioned reasons. Hence, in this paper, we introduce a method which can handle linear texts of arbitrary orientation, as well as curved text lines. In this paper we further propose HMM-based text verification for higher accuracy. There are two steps involved before the overlay text recognition is carried out, i.e., detection and extraction of overlay text. First, overlay text regions are roughly distinguished from background. The detected overlay text regions are refined to determine the accurate boundaries of overlay text strings. To generate a binary text image for video OCR, background pixels are removed from the overlay text strings in the extraction step. Although many methods have been proposed to detect and extract the video text, few methods can effectively deal with different color, shape, and multilingual text. Most of existing video text detection methods have been proposed on the basis of color, edge, and texture-based feature. Color-based approaches assume that the video text is composed of a uniform color. In the approach by Agnihotri et al. [4], the red color component is used to obtain high contrast edges between text and background. In [5], the "uniform color" blocks within the high contrast video frames are selected to correctly extract text regions. Kim et al. [6] cluster colors based on Euclidean distance in the RGB space and use 64 clustered color channels for text detection. However, it is rarely true that the overlay text consists of a uniform color due to degradation resulting from compression coding and low contrast between text and background.

Edge-based approaches are also considered useful for overlay text detection since text regions contain rich edge information. The commonly adopted method is to apply an edge detector to the video frame and then identify regions with high edge density and strength. This method performs well if there is no complex background and it becomes less reliable as the scene contains more edges in the background. Lyu et al. [7] use a modified edge map with strength for text region detection and localize the detected text regions using coarse-to-fine projection. They also extract text strings based on local thresholding and inward filling. In [8], authors consider the strokes of text in horizontal, vertical, up-right, and up-left directions and generate the edge map along each direction. Then they combine statistical features and use k-means clustering to classify the image pixels into background and text candidates.

III. PROPOSED MODEL

In this section, we present the description of proposed approach for text detection and recognition using image processing approach for natural video scenes. The complete approach is divided into two main phases as: (a) text detection and (b) recognition.

A. Text Detection in Natural Scenes

First of all, we apply text detection approach for given video frame data. Since, text in video frames or images can provide efficient information about the sequence hence accurate text detection and recognition is highly recommended for various real-time application systems. Several techniques are introduced during last decade for text detection where feature extraction based techniques shows significant impact on the detection. A study presented in [25] suggested that the combined feature extraction techniques can be helpful for the detection or localization of the text region in natural scenes. Based on this assumption, here we focus Hybrid Feature Fusion Model (HFFM) for text detection. The complete process is divided into two main components where low and high level feature extraction is performed for natural color images.

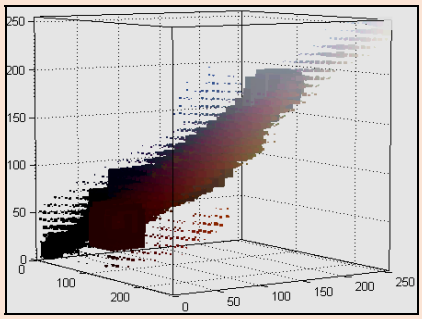
According to the proposed approach, we consider low level features such as SIFT and curvature related feature as discussed in our previous work [26]. We further enhance the low-level feature extraction model for text detection and categorized into three main categories which are based on the color continuity, color variation and intensity variation. Further, high level features are extracted and combined with the low-level feature.

B. Color Continuity

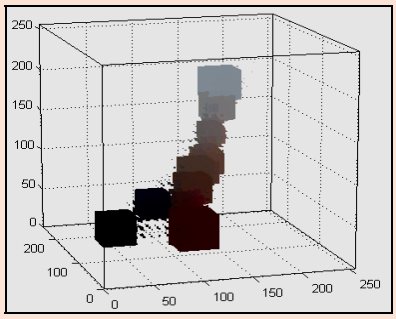
Generally, for image text extraction, regions need to be segmented for efficient analysis of the text region detection. Hence, in this work we apply color continuity estimation which can exploit the several regions of the color image. in order to achieve this, first of all we apply data sub sampling which reduces the number of samples in the data resulting in the reduced complexity and helps to obtain the significant color information.

In this process, first of all, binary edge operator using adaptive threshold method and a gradient for each color channel is computed which is used for magnitude construction for the input image.

(a) Input image



(b) Initial color distribution



(c) Color reduction distribution

Figure 2. Color Continuity Model

The magnitude computation can be expressed as:

$$M(x,y) = \max\left\{|\mathcal{G}^r(x,y)|, |\mathcal{G}^g(x,y)|, |\mathcal{G}^b(x,y)|\right\} \quad (1)$$

Where $\mathcal{G}$ denotes the gradient for different channels. The gradient of each channel can be expressed based on the number of rows, columns and pixel values. This can be computed as:

$$|\mathcal{G}^r(x,y)| = \sqrt{\left(\mathcal{G}^r_{row}(x,y)\right)^2 + \left(\left(\mathcal{G}^r_{col}(x,y)\right)^2\right)}$$
$$|\mathcal{G}^g(x,y)| = \sqrt{\left(\mathcal{G}^g_{row}(x,y)\right)^2 + \left(\left(\mathcal{G}^g_{col}(x,y)\right)^2\right)} \quad (2)$$
$$|\mathcal{G}^b(x,y)| = \sqrt{\left(\mathcal{G}^b_{row}(x,y)\right)^2 + \left(\left(\mathcal{G}^b_{col}(x,y)\right)^2\right)}$$

Let us consider that S number of samples are obtained using subsampling scheme with a 3D histogram. Here we assume that each sample is a candidate to the cluster center. The cluster candidate for each channel can be computed as:

$$R_r = \frac{1}{MN}\sum_M\sum_N I_R(x,y)$$
$$R_g = \frac{1}{MN}\sum_M\sum_N I_g(x,y) \quad (3)$$
$$R_b = \frac{1}{MN}\sum_M\sum_N I_b(x,y)$$

First of all, we apply, an initial color reduction where first of all $2h$ length cube is constructed. Let $s_1 = (r_1, g_1, b_1)$ denotes the center point of the cube, a new point $s_{m_1} = (r_{m_1}, g_{m_1}, b_{m_1})$ where r_{m_1}, g_{m_1} and , b_{m_1} denotes the values of red green and blue channels in the cube. These values can be expressed as:

$$r_{m_1} = \frac{\sum_{r=-h}^{h}\sum_{g=-h}^{h}\sum_{b=-h}^{h} r.\,M(r,g,b)}{\sum_{r=-h}^{h}\sum_{g=-h}^{h}\sum_{b=-h}^{h} M(r,g,b)}$$
$$g_{m_1} = \frac{\sum_{r=-h}^{h}\sum_{g=-h}^{h}\sum_{b=-h}^{h} g.\,M(r,g,b)}{\sum_{r=-h}^{h}\sum_{g=-h}^{h}\sum_{b=-h}^{h} M(r,g,b)} \quad (4)$$
$$b_{m_1} = \frac{\sum_{r=-h}^{h}\sum_{g=-h}^{h}\sum_{b=-h}^{h} b.\,M(r,g,b)}{\sum_{r=-h}^{h}\sum_{g=-h}^{h}\sum_{b=-h}^{h} M(r,g,b)}$$

In the next step, we estimate the available unique color in the image along with their probability of occurrence in the complete image data and arrange the obtained data into descending order of probability. Now, the data which has higher probability, will be added to the color palette and the center point and new bin values are updated using (4). Here, we compute Euclidean distance from the palette to the bin to identify the threshold for considering the bin for the threshold which is also helpful for selecting the color when there is no unique color is present. This process is repeated until the complete image is processed. Figure 1 shows a sample representation of initial color distribution and color distribution representation after reduction which is also considers the color clustering technique for the estimating the color continuity.

In the next phase, we apply multi-resolution scheme for edge detection which is useful for candidate point identification. The filtering using multi edge detector can be given as:

$$g(x,y) = \mathbb{D}.\,[\mathbb{G}(x,y)] * f(x,y) \quad (5)$$

Where $\mathbb{D}[.]$ denotes the derivative function, $\mathbb{G}$ represents the Gaussian function for (x,y) and $f(x,y)$ denotes the pixel density for the given input at pixel position (x,y). Once the edge filtering is done, we compute the threshold for image

using gray thresholding method and input image is binarized where initial components are extracted.

Color Variations

In this sub-section we present a brief discussion about proposed solution for the color variation-based approach for low-level feature extraction. Prior to this, we apply an initial process for the candidate region detection and the layout of the text which is present in the image. the text can be present in the any direction as it follows multi-orientation property. Hence, we present a color layout identification approach using initial color, position, size and text information obtained from the initial candidate component extraction. The color layout approach helps to achieve the optimal direction for processing. Initially, the search region can be given as:

$$\mathbb{D}(c, p, s, t_0) = \mathbb{D}0(c, p, s, t_0) \cup \mathbb{R}(c, p, s, t_0) \quad (6)$$

Here c, p, s and t_0 denotes the color, position, size and text which is identified initially, $\mathbb{D}0$ initial position of the direction movement, and $\mathbb{R}$ denotes the complete region. In order to further identify the accurate direction, color discrimination criteria by assuming that the similar texts will follow the same colors and the complete region where similar color is present, is analyzed and obtained color information related to the text is stored. In this work, we use Gaussian mixtures for color modeling for both text and background region. According to this, Gaussian mixture model, d dimensional color vectors are considered which are represented into K cluster form, the probability distribution of this data can be expressed as:

$$Pr(x) = \sum_{k=1}^{K} \psi_k(x).w_k \quad (7)$$

Where w_k denotes the mixture weight, and Gaussian function is used for basis function as:

$$\psi(x) = \frac{1}{\left(|C|^{\frac{1}{2}}(2\pi)^{\frac{D}{2}}\right)} \exp\left\{-\frac{1}{2}((x-m)^T)((x-m)^T C^{-1})\right\}$$

where m_x and C represents the covariance matrix. With the help of these assumptions, the probability of pixel generation with color vector can be computed as:

$$\lambda_k(x) = \frac{\psi_k(x).w_k}{p(x)} \quad (8)$$

Here, we apply, EM approach for optimal parameters of C and w for the given K number of Gaussian mixtures, the EM likelihood maximization can be given as:

$$L = \prod_{n=1}^{N} p(x) \quad (9)$$

Where N denotes the total number of pixels identified in the region obtained from the color layout search.

D. Text Localization

Here, first of all we extract the optimal text region for the given input binary image whose initial candidates are extracted using image binarization approach. Let us consider that the T denotes the set of pixels in the given image I where we need to perform the grouping of text-like region using a probabilistic methods as $Pr(s|S, I)$ where higher probability regions are clustered together to form a text region. Hence, we apply vertical and horizontal edge mapping $\mathbb{C}_v$ and $\mathbb{C}_h$which are obtained using directional Canny filtering. Later, based on the types of edges, horizontal and vertical directional edges are extended and can be represented as:

$$\mathbb{D}_v(s) = \mathbb{C}_v(s) \oplus R_v \text{ and } \mathbb{D}_h(s) = \mathbb{C}_h(s) \oplus R_h \quad (10)$$

Where R_v and R_h denotes the vertical and horizontal shapes which are represented in the rectangle of 1×5 and 6×3 respectively. The sample outcome of vertical and horizontal shape is depicted in figure 3.

(a) **(b)** **(c)**

Figure 3. (a) input image (b) vertical direction shape and (c) horizontal shape.

The character strokes and vertical edges are related to each other which are generally connected to the horizontal edges also if any text region is present, vice versa. Here, we consider only the region which is covered using vertical and horizontal dilation of filtered edges. Hence, the probability can be given as $P(s|S,I) = D_v(s)D_h(s)$

After identification of text region, we perform the text verification whether it belongs to the text or not using baseline methods. The proposed approach helps to remove the various complexities such as skew correction, false positives etc. With the help of this process of baseline, Y direction projection $h(y)$ is initiated where y denotes the total number of pixels in y. Here, we introduce a constant called as density factor which provides the information about available pixels in the detected text region boundaries. Here, we iteratively split the horizontal lines based on the obtained threshold. This splitting can be performed using three basic stages:

1. Splitting of equal length: in this approach, two text line of similar length are analyzed which can be split using Otsu's method of thresholding. Let $h(y)$ denotes a one-dimensional histogram where Otsu thresholding is applied to minimize the intra-class difference between the detected lines. At this stage, if the length of current obtained line $h(y_0)$ is less than the 50% of the longest line, then we split the region at y_0
2. Varying length splitting: in this stage, we consider the text lines which contains different length of text strings or the text strings which are connected to the background.
3. Baseline refinement: if any detected text region is not identified under the above mentioned technique, then the complete identified region is segmented.

With the help of these methods, the false positive can be removed which helps to improve the performance of the system.

E. Feature Extraction (Text Verification and Recognition)

Once the text lines are detected and localized, we apply recognition strategy which is used for text recognition from the natural scene images. In this work, we compute several features such as grayscale spatial derivative which are obtained by computing the intensity in both the X and Y direction which results in to a feature vector, distance map feature, gradient and DWT features.

F. Distance Map Based Robust Feature

In this sub section, we present distance map feature modeling where it is assumed that the text character contrast depends on the background data where background brightness spatial derivatives may vary according to the textcharacteristics.in order to maintain the robustness of the feature vector we consider a distance map-based feature generation process where key features are extracted from the center of the edge image and the obtained feature map can be defined as:

$$\forall_s \in S, D_M(s) = \min_{s_1 \in E} d(s, s_1) \tag{11}$$

Where d denotes the Euclidean distance function, pixel and edge sets are related as $E \subseteq S$. In order to further improve the obtained distance feature, we incorporate gradient feature vector which is used for normalizing the contrast at a given point and its neighboring point. Let $g(s)$ denotes the magnitude of the gradient at s and local mean is denoted as $LM(s)$, hence the local mean and variance can be defined as:

$$LM(s) = \frac{1}{|G_s|}\sum_{s_1 \in G_s} g(s_1) \quad \text{and} \quad LV(s) = \frac{1}{|G_s|}\sum_{s_1 \in G_s}(g(s_1) - LM(s))^2 \tag{12}$$

Then, finally it can be expressed as:

$$G(s) = \sqrt{\left(\frac{GV}{LV(s)}\right)}\,(g(s) - LM(s)) \quad (13)$$

Where GVrepresents the global gradient variance computed over the whole image grid S.

Further, we apply DWT (Discrete wavelet Transformation) based feature extraction modeling for to obtain the frequency domain feature which are further used for the texture analysis of the given image. figure 2 presents a DWT based feature model for the processed image.

F. Classifier Construction

After feature extraction process, we model these features in to a machine learning process where the generated features of the database are learned using statistical learning technique. In this work, we apply SVM (Support Vector Machine) classifier. Support vector machine shows a significant impact on the high dimensional dataspaces such as obtained texture patterns. According to the key idea of SVM, the input data of pattern is projected on the high-dimensional space which is also called as feature space. With the help of SVM, the projected data can be divided into two or more class-labels which can be separated linearly. Let us consider that n number of training patterns are present as $(x_1, y_1), \ldots, (x_n, y_n)$ where $y \pm 1$ represents the negative or positive classes based on the hyperplane expressed as $w.\phi(x) + b = 0$ in the feature space which contains two separable classes. In this process, the training samples can be expressed as $w = \sum_i \gamma_i y_i \phi(x_i)$where $\gamma_i \geq 0$. Here, SVM classifier need to be trained hence we used standard quadratic programming technique using a Radial basis function (RBF) which is expressed as:

$$K(x, x_i) = e^{-\frac{\|x - x_i\|}{2\sigma^2}} \quad (14)$$

Where σ denotes the kernel size(which is obtained using M-fold cross validation) ,based on these functions, the SVM classification model can be presented as:

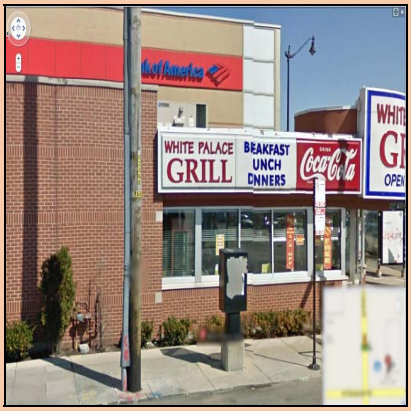
$$G(z) = \sum_{i}^{n} \gamma_i y_i \phi(x_i).\phi(z) + b = \sum_{i}^{n} \gamma_i y_i K(x_i, z) + b \quad (15)$$

V. RESULTS AND DISCUSSION

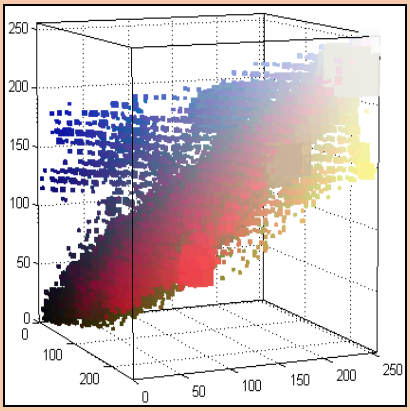
This section presents a complete experimental study for text recognition system. This process is implemented using MATLAB tool on windows platform. In this experiment, we have considered three standard databases for image text recognition which are called as: The Street View Text (SVT) [27], ICDAR 2003 (IC03) [28] and IIIT 5K-word (IIIT5K) [29]. The SVT dataset contains total 647 images collected from the Google street view of the road side scenes. This dataset provides 50 words lexicon for each image hence it is also denoted as SVT-50. The ICDAR 2003 dataset contains total 860 words which are obtained from the 251 natural images by cropping the words. This dataset provides 50 lexicons and full dictionary hence it is called as IC03-50 and IC03-FULL, respectively. The IIIT5K dataset contains total 5000 images where both digital and scene images are present.

The complete process of text detection and recognition is depicted in figure 4 where all stages of proposed approach are presented.

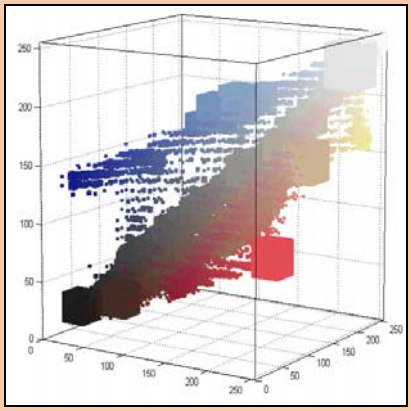
.



(a) Input image

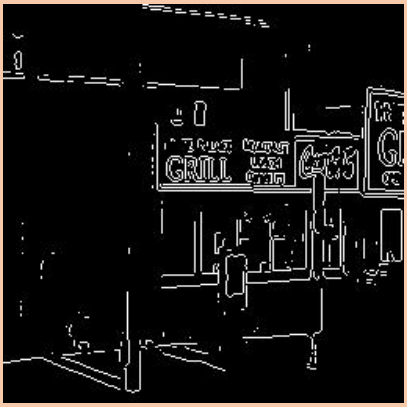


(b) Initial color distribution

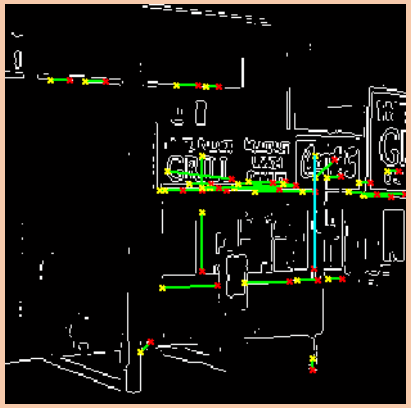


(c) Color reduction distribution

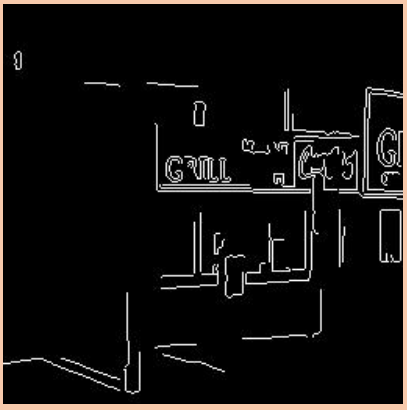
(d) Candidadate point extraction



(e) Edge detection

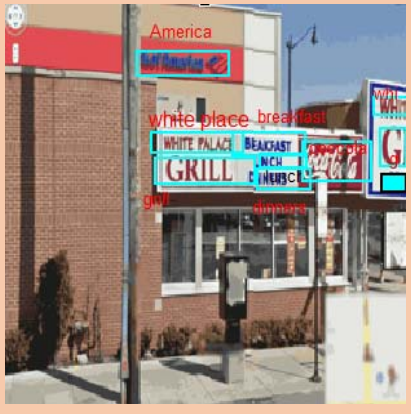


(f) Line detection



(g) Line removal

(h) Text detection



(i) Recognized text

Figure 4. Complete illustration of proposed results

Based on the proposed approach, we present a comparative study where we have considered the above mentioned datasets and compared the performance in terms of recognition accuracy.

TABLE I.RECOGNITION ACCURACY PERFORMANCE

Technique	ICDAR 2003	IC03-Full	SVT-50	IIIT5k-50	IIIT5k-1k
Wang et al. [27]	76	62	57	64	57
Mishra et al [29]	81.8	67.8	73.2	-	-
TSM+CRF et al. [30]	87.4	79.3	73.5	-	-
Lee et al. [31]	88	76	80	-	-
Wang et al. [32]	90	84	70	-	-
Alsharif et al. [33]	93.1	88.6	74.3	-	-
Su et al. [34]	92	82	83	-	-
Almazan et al. [35]	-	-	87	88.6	75.6
Photoocr [36]	-	-	90.4	-	-
Jaderberg et al. [37]	97.8	97	93.2	95.5	89.6
DTRN [39]					
Jaderberg et al. [38]	98.7	98.6	95.4	97.1	92.7
Proposed Approach	98.79	98.55	95.26	97.5	95.8

Table I shows a comparative analysis for different datasets. This study shows that proposed approach achieves better performance when compared with the existing techniques.

Similarly, we have considered ICDAR 2015 video database where 25 videos are providing for training and 24 videos are provided for testing purpose. The performance of proposed approach is evaluated and compared with existing techniques in terms of precision, recall and F-score. The comparative performance is presented in table II.

TABLE II. COMPARATIVE PERFORMANCE FOR THE VIDEO TEXTLOCALIZATION USING ICDAR-2015DATABASE

Technique	Precision	Recall	F-Score
CNN MSER [41]	0.3471	0.3442	0.3457
Deep2Text-MO [42]	0.4959	0.3211	0.3898
AJOU [43]	0.4726	0.4694	0.4710
NJU [41]	0.7044	0.3625	0.4787
StradVision1 [41]	0.5339	0.4627	0.4957
StradVision2 [41]	0.7746	0.3674	0.4984
Zhang et al. [48]	0.708	0.4309	0.5358
Tian et al. [34]	0.7422	0.5156	0.6085
Yao et al. [41]	0.7226	0.5869	0.6477
PVANET2x RBOX MS [40]	0.8327	0.7833	0.8072
PVANET2x RBOX [40]	0.8357	0.7347	0.7820
PVANET2x QUAD[40]	0.8018	0.7419	0.7707
VGG16 RBOX [40]	0.8046	0.7275	0.7641
PVANET RBOX[40]	0.8063	0.7135	0.7571
PVANET QUAD[40]	0.8119	0.6856	0.7434
VGG16 QUAD[40]	0.7987	0.6895	0.7401
Proposed approach	0.8566	0.9126	0.8213

This study shows that the proposed approach achieves better performance when compared with the existing techniques.

V. CONCLUSION

In this work, we have focused on the development of text detection and recognition technique for video scene texts. The proposed method contains several stages to obtain the text recognition performance where low level and high level features are extracted and combined to distinguish between foreground and background region. In order to achieve the low level feature representation, we consider SIFT features as the base feature and later color continuation, color variation, and intensity variation features are extracted along with the edge refinement techniques to achieve the information about the text region. In the next phase, baseline method is applied to estimate the lines in the binarized image and text localization is performed based on the vertical and horizontal shape projections. Finally, feature extraction and learning process is applied for text recognition where distance map feature, DWT features are extracted and database is trained using SVM classifier. An extensive experimental study is carried out on the different type of datasets and compared with the state-of-art techniques. The comparative shows that the proposed approach achieve better performance when compared with the existing techniques.

REFERENCES

[1] Ye, Q. and Doermann, D.," Text detection and recognition in imagery: A survey", IEEE transactions on pattern analysis and machine intelligence, 37(7), pp.1480-1500, 2015.

[2] Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R., Darrell, T. and Saenko, K., "Sequence to sequence-video to text", In Proceedings of the IEEE international conference on computer vision pp. 4534-4542, 2015.

[3] Yin, X.C., Zuo, Z.Y., Tian, S. and Liu, C.L., "Text detection, tracking and recognition in video: a comprehensive survey", IEEE Transactions on Image Processing, 25(6), pp.2752-2773, 2016.

[4] Wu, L., Shivakumara, P., Lu, T. and Tan, C.L., "A New Technique for Multi-Oriented Scene Text Line Detection and Tracking in Video", IEEE Trans. Multimedia, 17(8), pp.1137-1152, 2015.

[5] Tang, X., Gao, X., Liu, J. and Zhang, H., "A spatial-temporal approach for video caption detection and recognition", IEEE transactions on neural networks, 13(4), pp.961-971, 2002.

[6] J. Zang and R. Kasturi, "Extraction of Text Objects in Video Documents: Recent Progress", In Proceedings of Document Analysis Systems (DAS), pp. 5-17, 2008.

[7] Lienhart, R., "Video ocr: A survey and practitioner's guide", In Video mining, Springer, Boston, MA, pp. 155-183 2003.

[8] Yi, C. and Tian, Y., 2011. Text string detection from natural scenes by structure-based partition and grouping. IEEE Transactions on Image Processing, 20(9), pp.2594-2605.

[9] J. Zang and R. Kasturi, "Extraction of Text Objects in Video Documents: Recent Progress," In Proceedings of Document Analysis Systems (DAS), pp. 5-17, 2008.

[10] Khan, J.F., Bhuiyan, S.M. and Adhami, R.R., 2011. Image segmentation and shape analysis for road-sign detection. IEEE Transactions on Intelligent Transportation Systems, 12(1), pp.83-96.

[11] Yu, M. and Kim, Y.D., 2000. An approach to Korean license plate recognition based on vertical edge matching. In Systems, Man, and

Cybernetics, 2000 IEEE International Conference on (Vol. 4, pp. 2975-2980). IEEE.

[12] Ye, Q. and Doermann, D., 2015. Text detection and recognition in imagery: A survey. IEEE transactions on pattern analysis and machine intelligence, 37(7), pp.1480-1500.

[13] Lu, W., Sun, H., Chu, J., Huang, X. and Yu, J., 2018. A Novel Approach for Video Text Detection and Recognition Based on a Corner Response Feature Map and Transferred Deep Convolutional Neural Network. IEEE Access, 6, pp.40198-40211.

[14] Roy, S., Shivakumara, P., Jain, N., Khare, V., Dutta, A., Pal, U. and Lu, T., 2018. Rough-fuzzy based scene categorization for text detection and recognition in video. Pattern Recognition, 80, pp.64-82.

[15] Roy, S., Shivakumara, P., Roy, P.P., Pal, U., Tan, C.L. and Lu, T., 2015. Bayesian classifier for multi-oriented video text recognition system. Expert Systems with Applications, 42(13), pp.5554-5566.

[16] Huang, X. and Ma, H., 2010, August. Automatic detection and localization of natural scene text in video. In Pattern Recognition (ICPR), 2010 20th International Conference on (pp. 3216-3219). IEEE.

[17] Shivakumara, P., Phan, T.Q. and Tan, C.L., 2011. A laplacian approach to multi-oriented text detection in video. IEEE transactions on pattern analysis and machine intelligence, 33(2), pp.412-419.

[18] Li, Y. and Lu, H., 2012, November. Scene text detection via stroke width. In Pattern Recognition (ICPR), 2012 21st International Conference on (pp. 681-684). IEEE.

[19] Shi, C., Wang, C., Xiao, B., Zhang, Y. and Gao, S., 2013. Scene text detection using graph model built upon maximally stable extremal regions. Pattern recognition letters, 34(2), pp.107-116.

[20] Yang, H., Quehl, B. and Sack, H., 2014. A framework for improved video text detection and recognition. Multimedia Tools and Applications, 69(1), pp.217-245.

[21] Liang, G., Shivakumara, P., Lu, T. and Tan, C.L., 2015. Multi-spectral fusion based approach for arbitrarily oriented scene text detection in video images. IEEE Transactions on Image Processing, 24(11), pp.4488-4501.

[22] Khare, V., Shivakumara, P., Raveendran, P. and Blumenstein, M., 2016. A blind deconvolution model for scene text detection and recognition in video. Pattern Recognition, 54, pp.128-148.

[23] Mittal, A., Roy, P.P., Singh, P. and Raman, B., 2017. Rotation and script independent text detection from video frames using sub pixel mapping. Journal of Visual Communication and Image Representation, 46, pp.187-198.

[24] Bhunia, A.K., Kumar, G., Roy, P.P., Balasubramanian, R. and Pal, U., 2018. Text recognition in scene image and video frame using Color Channel selection. Multimedia Tools and Applications, 77(7), pp.8551-8578.

[25] Ji, Z., Wang, J. and Su, Y.T., 2009, July. Text detection in video frames using hybrid features. In Machine Learning and Cybernetics, 2009 International Conference on (Vol. 1, pp. 318-322). IEEE.

[26] Segu, R. and Suresh, K., 2017. Joint feature extraction technique for text detection from natural scene image. International Journal of Signal and Imaging Systems Engineering, 10(1-2), pp.14-21.

[27] Wang, K.; Babenko, B.; and Belongie, S. 2011. End-to-end scene text recognition. IEEE International Conference on Computer Vision (ICCV)

[28] Lucas, S. M.; Panaretos, A.; Sosa, L.; Tang, A.; Wong, S.; and Young, R. 2003. Icdar 2003 robust reading competitions. International Conference on Document Analysis and Recognition (ICDAR).

[29] Mishra., A.; Alahari, K.; and Jawahar, C. 2012. Scene text recognition using higher order language priors. British Machine Vision Conference (BMVC).

[30] Shi, C.; Wang, C.; Xiao, B.; Zhang, Y.; Gao, S.; and Zhang, Z. 2013. Scene text recognition using part-based treestructured character detection. IEEE Computer Vision and Pattern Recognition (CVPR).

[31] Lee, C.; Bhardwaj, A.; Di, W.; and andR. Piramuthu, V. J. 2014. Region-based discriminative feature pooling for scene text recognition. IEEE Computer Vision and Pattern Recognition (CVPR).

[32] Wang, T.; Wu, D.; Coates, A.; and Ng, A. Y. 2012. End-to-end text recognition with convolutional neuralnetworks. IEEE International Conference on Pattern Recognition (ICPR).

[33] Alsharif, O., and Pineau, J. 2013. End-to-end text recognition with hybrid HMM maxout models. arXiv:1310.1811v1.

[34] Su, B., and Lu, S. 2014. Accurate scene text recognition based on recurrent neural network. Asian Conference on Computer Vision (ICCV)

[35] Almazan, J.; Gordo, A.; Forn ´ es, A.; and Valveny, E. 2014. ´ Word spotting and recognition with embedded attributes. IEEE Trans. Pattern Analysis and Machine Intelligence (TPAMI) 36:2552–2566.

[36] Bissacco, A.; Cummins, M.; Netzer, Y.; and Neven, H. 2013. Photoocr: Reading text in uncontrolled conditions. IEEE International Conference on Computer Vision (ICCV)

[37] Jaderberg, M.; Simonyan, K.; Vedaldi, A.; and Zisserman, A. 2015a. Deep structured output learning for unconstrained text recognition. International Conference on Learning Representation (ICLR).

[38] Jaderberg, M.; Simonyan, K.; Vedaldi, A.; and Zisserman, A. 2015b. Reading text in the wild with convolutional neural networks. International Jounal of Computver Vision (IJCV).

[39] He, P., Huang, W., Qiao, Y., Loy, C.C. and Tang, X., 2016, February. Reading Scene Text in Deep Convolutional Sequences. In AAAI (Vol. 16, pp. 3501-3508).

[40] Zhou, X., Yao, C., Wen, H., Wang, Y., Zhou, S., He, W. and Liang, J., 2017, July. EAST: an efficient and accurate scene text detector. In Proc. CVPR (pp. 2642-2651).

[41] D. Karatzas, L. Gomez-Bigorda, A. Nicolaou, S. Ghosh, A. Bagdanov, M. Iwamura, J. Matas, L. Neumann, V. R. Chandrasekhar, S. Lu, F. Shafait, S. Uchida, and E. Valveny. ICDAR 2015 competition on robust reading. In Proc. of ICDAR, 2015.

[42] X. C. Yin, W. Y. Pei, J. Zhang, and H. W. Hao. Multiorientation scene text detection with adaptive clustering. IEEE Trans. on PAMI, 37(9):1930–1937, 2015.

[43] H. Koo and D. H. Kim. Scene text detection via connected component clustering and nontext filtering. IEEE Trans. On Image Processing, 22(6):2296–2305, 2013.