Text Spotting in Video: Recent Progress and Future Trends

Manasa Devi Mortha, Seetha Maddala, Viswanadha Raju Somalaraju

Abstract— The wid¬¬e popularity of videos, images, documents available in the internet have led to the demand for automatic annotations, indexing, construction of videos and many other applications. In order to implement these demands, text is a major cause of information which requires detection, localization, tracking and recognition process. Nevertheless, text variation owing to font-size, font-style, direction, occlusion, poor resolution makes automatic text extraction more challenging. Thus, video pre-processing plays a vital role before detecting and recognizing the text. This paper emphasizes on survey for different detection and recognition methods, feature descriptors, datasets, and performance and evaluation process from diverse publications. Traditional methods like connected components, region based, texture based, Neural Networks, OCR have been reviewed. Among which Scale Invariant Feature Transform (SIFT), Maximally Stable Extremal Regions (MSER), Convolution Neural Networks (CNN) are effective and efficient feature descriptors in spotting the text. However, this paper shows comparative study of ubiquitous features descriptors along with their dependant parameters which declines the performance of recognizing the video text. Conversely, hybrid methods and CNN techniques have done magnificent work to achieve text spotting in scene images on different datasets like ICDAR, ImageNet, and CIFAR10 etc. However, ICDAR 2013/15 is specially prepared to challenge the detection of text in videos. Finally, related performance metrics and future trends for video text spotting are comprehensively analysed.

Keywords: Recognition, text detection, caption text, tracking, natural scene text, Convolution NN, video pre-processing, feature descriptors

1. INTRODUCTION

With the increase in adulation of smart phones, computers, other smart machines which allow uploading billions of videos and hours and hours of streaming, made people to watch lots of videos of their interest. With this, finding the appropriate video of interest becomes a challenging research work. Traditionally, manual tagging was used to annotate the video for searching and indexing. This traditional approach could not able to find the video significantly and it is time taking process to tags. Many research was went on this, to find the state-of-the-art model to find the text from video for tagging purpose. Text in image or video comprises of crucial data and utilized in numerous content-based imagery and visual applications. Text varies in font, size, orientation, lighting, background, texture, which makes more difficult to detect in video. Literature divides text into duplet, graphics text and scene text. Graphics text is also named as artificial text or caption text which included in the video. Generally, many news channels are having scrolling of text, which is added to the natural scene. Scene text is the natural text which comes with the scenes, like number plates, house numbers, boarding's, text on the shirt, etc.; Detecting caption text is not difficult as it will be in same font size, lighting, texture etc.; but detecting scene text is a challenging with varying orientation, font, color. Many studies have done on detecting caption text in natural scene imageries and video, but very rare work has done on both the texts. [1] Comprehensively contemplated a numerous technique for detecting and extricating the text from videos or images. [2] Delivered the structure for detecting multi-oriented text from video by using spatio-temporal data, respectively. [3] Provided a robust method for detecting text. They have used MSER pruning to extract character image patches which are collected into candidate text by single link clustering algorithm. Although many text detection methods are proposed and implemented for indexing and annotations, there are many other applications like supermarket automation, merchandise movement, licence plate recognition etc. These are useful for the society in terms of efficiency and speed in processing the things. Text Detection gives raise to many applications, but video-based text detection is a challenging task where it requires following stages: text tracking, detection, localization, and recognition. In the literature, many methodologies have been proposed in terms of text localisation, which targets to determine the position of the text. The process of detection is utilized to decide whether any text is present or not. Text recognition is used to classify the text from non-text. Enhancement process can be encompassed to improve the resolution prior to recognition. However, applying fusion techniques to improve the resolution prior to recognition will lead to high efficiency and accuracy. Previous work on text detection in video using fusion techniques had been proposed, but still need improvement which increases the recall and precision. [18] Proposed an innovative approach which combines Laplace operator by high frequency high band wavelet across multi-level fusion to recognize candidate text. They have also used MSER with SWT to preserve fine details of text candidates. Once images are fused, they are fed to machine to classify the text in the appropriate group. Commercial OCR products are successful in detecting the text from document image but not well on video text. The principal goal of text

Revised Version Manuscript Received on April 05, 2019.

Manasa Devi Mortha, VNR Vignana Jyothi Institute of Engineering & Technology, Department of Computer Science & Engineering, Vigana Jyothi Nagar, Hyderabad, India.(E-Mail: manasadevi_m@vnrvjiet.in)

Dr. Seetha Maddala, G.Narayanamma Institute of Technology & Science, Department of Computer Science & Engineering, Shaikpet, Hyderabad, India.(E-Mail: seetha.maddala@gmail.com)

Dr. Viswanadha Raju Somalaraju, Department of Computer Science & Engineering, Jawaharlal Nehru Technological University – Jagtial, Nachupally, Jagtial, India.(E-Mail: svraju.jntu@gmail.com)

discovery, detection and recognition is indispensable for an endwise system.

Fig. 1 Graphics/Caption/Artificial Text Example

Fig. 1 Scene Text Example

2. VIDEO PRE-PROCESSING

Text detection in video is one among the leading zones under pattern recognition. Text discovery system is not easy to design because of variation in font size, style, color, orientation, textured background with low contrast. Thus, we need video pre-processing to reduce complexity of detecting, localizing, extracting, tracking and recognizing of video text. Following figure demonstrates a typical pre-processing phase in video text detection [47]. Fig. 2.1a is an initial video frame employed by image segmentation process where all the pixels have been extracted which is appropriate towards text (Fig. 2.1b). Normally, the aim of pre-processing operations for text detection from video is to improve the imagery data by vanquishing undesired degradations besides improving text applicable features. There are few pre-processing operators to enhance features or supress information which is not significant to visual text detection.

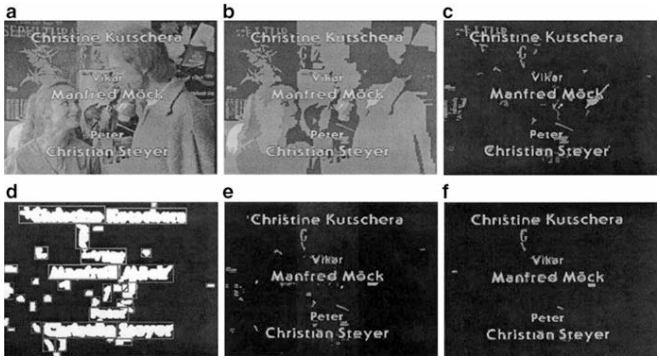


Fig.2.1. Preliminary states of video text detection. (a) initial visual frame, (b) dissection phase, (c) dimension restraint phase, (d) binary process and dilation, (e) movement analysis, and (f) result of dissimilarity analysis [31]

a. Image Cropping

In this, inappropriate parts in the imagery are cropped so that supplementary processing emphases on interest regions and thereby computational cost are reduced. Local operator [31] is a kind of image processing transformations. In this process, value of individual pixel of output depends only on corresponding pixel value of an input. A native operator is denoted by,

$$\mathbf{m(d) = n\ (G\ (e))}$$

'd' is a location of an imagery pixel and G(e) is pixel value. There are many other used operators [60, 24], where relevant statistics like minimal, maximal and the average values of intensity is calculated to crop the relevant image region.

b. Neighbourhood operators

It is used to remove noises and sharps the image details. Linear filter is the most frequently used operator in which the value of pixel of output is a linear arrangement of the values of the input pixel's neighbourhood.

$$\mathbf{N\ (p, q) = \sum F\ (p - r, q - s) * m\ (r, s)}$$

Whereas, m (r, s) is known as weight mask, r, s defines range of neighbours. [47] Describes about the types of filters available to smooth the image. Average filter is a simplest filter which performs mean on all the values of the pixel in a C*C window. It is also used to reduce the distortion or blur. There is another filter called as Median filter used for smoothing. It picks the middle value from the neighbourhood as the outcome for each pixel. The main intent of sharpening is to high-point the information in a frame. Below is the difference of the equation for 1st order derivative of 1D function f'(x).

$$\partial f'/\partial x = f'(x+1) - f'(x)$$

Roberts proposed commonly used differential operator [61] as

$$\mathbf{G_x = (z_9 - z_5)}$$
$$\mathbf{G_y = (z_8 - z_6)}$$

It is calculated as,

$$\mathbf{\nabla f = |z_9 - z_5| + |z_8 - z_6|,}$$

Which is a Gradient Approximation

Another frequently used operator is Sobel operator for first-order [25]. It uses the values of absolute at point z_5 with a kernel of 3*3 for gradient approximation. It follows as:

$$\nabla f = |(z_7+2z_8+z_9) - (z_1+2z_2+z_3)| + |(z_3+2z_6+z_9) - (z_1+2z_4+z_7)|$$

For video frame processing, the difference between pixels can be calculated as

$$\partial f'2/\partial x2 = f'(x+1) + f'(x-1) - 2f'(x)$$

called the **Laplacian filter**. Below are the outcomes of all three types of operators shown in Fig. 2.2, respectively.

$$\begin{bmatrix} -1 & 0 & +1 \\ -2 & 0 & +2 \\ -1 & 0 & +1 \end{bmatrix} \quad \begin{bmatrix} +1 & +2 & +1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix}$$

Sobel Operates

$$\begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix} \quad \begin{bmatrix} 1 & 1 & 1 \\ 1 & -8 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

Laplacian Operators

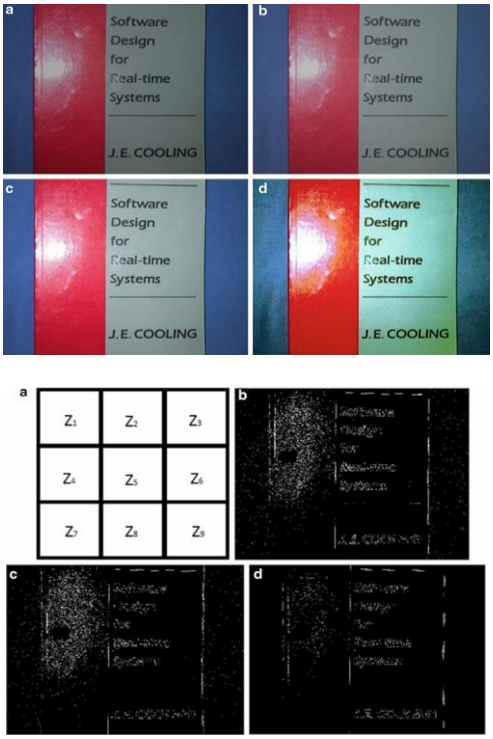


Fig. 2.2 filters samples. (a) Table which demonstrates filters, (b) Roberts filter, (c) Sobel filter, and (d) Laplacian filter [31]

c. Other operators

There are other operators like Morphological [26] which facilitates the video text detection or OCR by changing the nature of an underlying object. Color-based pre-processing mainly focus on conversion of RGB to grayscale representation.

$$\text{gray} = \max(R, G, B) + \min(R, G, B) / 2$$

Hasan and Karam [25] presented the a color input image by RGB components for text extraction.

$$\text{gray} = 0.299\,R + 0.587\,G + 0.114\,B$$

After color transformation, few regions of text which are eminent in color imagery come to be challenging to be detected in gray-level imagery. In order to resolve this problematic issue, [26] used homogeneous intensity of text areas under imageries. Texture analysis observes that text regions commonly have prominent Textura attributes from their backgrounds. Accordingly, spatial variance [27] along each parallel line is used to discriminate the background from text areas.

d. Motion Analysis

As a pre-processing procedure, motion vector analysis is believed to support for text detection in visual from temporal frames. [48] Present a solution for robotic text extraction for digital audio-visual build on movement analysis. [29] Used optical flow technique for motion analysis to exploit the temporal redundancy for text extraction for the same frame.

3. FUNDAMENTAL APPROACHES FOR TEXT SPOTTING

3.1 Text Detection Techniques

The objective of text detection is to localize the text in each video frame and draw a bounding box around the text. It localizes text components and clusters into candidate text regions.

3.1.1 Methods

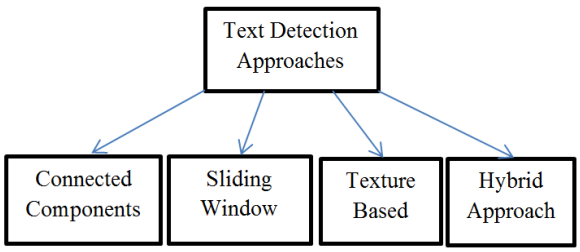
Prevailing approaches for text detection from scenes can be classified into following: Connected components, sliding region-based techniques and hybrid methods.

Connected Component (CC) based systems extract character candidates from video frame by connected component analysis followed by assembling candidates into text. They are good for imageries with great contrast and plain background. Text in video cannot be detected with CC because of poor resolution and reformatted background. This can be applied on caption text. The features used are color, edges, strokes and texture.

Sliding window-based approach is also termed as region-based techniques. It slides the small window or kernel or mask on video frame to search for possible text. Later machine learning procedures are applied to detect texts. These procedures are time consuming as the imagery must get processed in diverse scales. It utilizes many features like edge, texture, gradients and other correlated region characteristics to discriminate between text and non-text. These approaches are fast, overcome poor contrast problems, however, when the background is complex, they produce many false positives.

Texture based methods work well for complex backgrounds. It uses several methods like Gabor filter, FFT, spatio-variance, WT (wavelet transforms) or process of multiple channels for computing the chunks with texture. Then and there, the chunks are fed to the classifier to categorise text from non-text. These methods tend to be unsuccessful when text-like texture appears in the background.

Hybrid method makes use of two or methods to detect text, like Phan exploits region sensor to spot text and removes characters by connected graph components by local binarization process, non-characters are abolished with CRFs model. [4] Proposed object recognition-based methods which extracts features from the region and then use convolution neural network classifier for recognition of text. [5] Provided a novel framework for detecting text by taking benefit of both sliding window based and MSERs methods. Moreover, convolution NN is pragmatic to separate the networks of various characters components.



Among all the above-mentioned methods, MSER and SWT are magnificent in performance. MSER dependent methods maintain uniformity of text strokes by utilizing color (intensity) information. SWT employs the homogeneity of

text strokes thickness to detect the text. But SWT can detect horizontal text, gives more false positives. Non-text will not be detected, whereas MSER provides robustness to symmetrical and radiance conditions.

[31, 32] proposed combination of Laplace operator and Fourier transform approach for multi-oriented text detection from video to enhance low contrast texts and K-means for obtaining text candidates. Concept of skeleton is utilized for segmentation of text outlines grounded on final and intersection points. Still, this tactic gives low precision due to determination of exact end and junction points. Li et al. suggested the amalgamation of wavelet moments. This approach is noble for horizontal text.

[46] Contributed architecture for detecting and classifying the text into character case-sensitive, insensitive and bigram with convolution neural network classifier. They also used mining to mine and annotate the text and later CNN for classification, which slowed down the detection process. [6] Projected a technique to detect text based on character and link energies. Every two adjacent characters, in this approach, are coupled by a linkage. It is called as text unit. But this process is not suitable for languages in which several characters in a word are not associated together. [7] Surveyed many techniques and analysed, compared scientific challenges, procedures and the performance of text discovery and recognition in colour picture. [8] Provided system with endwise, which associates lexicon using modules of discovery or recognition by post processing methods together with non-maximal suppression and beam search.

[9] Surveyed many papers and classified into methods based on edge, texture, connected components, stroke width transforms, MSER for text detection and localization and algorithms for text enhancement as well as segmentation. They have also mentioned that there is no unique method which can satisfy all the cases, because, scene text varies in typefaces, radiance, haze, distortions. [11] Used fast and robust text detection method which finds all possible text lines in an image. They have used wavelet feature in locating candidate pixels and region growing system to link candidate pixels to form a region. [12] Proposed a fusion based text detection and localization for video image where it involves fusion of various edge-based method which works in bottom up fashion. This proposed method does not work on complex background.

[13] Used an Extremal Regions (ERs) for detecting a character in a scene image. ERs are sturdy to illumination, blur, texture variation, color and poor contrast. Robustness of this method is validated by false positives triggered through watermark text in the dataset against noise and poor contrast character.

[14] Proposed a shot segmentation for video indexing. They have divided text extraction problem into three key tasks- text discovery, localization and subdivision. This proposed algorithm, and maximum stroke width and minimum difference with background cannot detect very small fonts.

3.1.2 Attributes

Color features. Generally, text in video will be in consistent color unless it is scene text. Because scene text comes with the video, whereas, artificial text is added. It is easy to spot artificial text with similar color feature and contrast with the background. This feature is used to localize text [36, 40,61]. It is efficient in localizing the text even though the color is sensitive to multi-colored typescripts and irregular illumination. [52] Used clustering algorithm to attain connected components with similar color. [56] Obtained layers of color to progress the complex background for robustness by using mean-shift algorithm. [39] Used k-means algorithm to perform text extraction in the HSV color space. [41] Proposed Mixture of Gaussians model in RGB, intensity and Hue networks for text localization.

Structural features. Edges are reliable feature for text detection. In [19], edge is used for comparison between frames for similarity check. The frames are similar when the inter frame space difference is high and stroke for one frame and reject all the remaining frames.

Textura features. Dense characters are considered as texture. It includes Fourier Transformations, DCT, Wavelet, and HOG have been used to localize text. Texture attributes are effective in spotting dense characters.

Stroke Width Transform. It is a resident imagery operator that calculates the breadth of the most probable stroke containing pixel. It produces a map, where each component resembles to the pixel's stroke width value. In [8], strokes are considered as a characteristic, because text and non-text contains strokes with approximately constant thickness which results two parallel sets of edge in their frontiers. Similarly, in [4], gradient direction is used to regularize the input frame's edge map, causing in removing the undesirable elements.

MSERs.It has been widely explored in [5, 6]. It has been perceived that text elements usually have substantial color disparity with backgrounds and incline to form homogenous color areas. This algorithm which adaptively spots stable color areas provides a feasible result for localizing text. [6], Contributed model of deep learning to challenge two key issues of existing MSERs techniques for text detection, enables the structure with strong robustness and high discriminative ability to differentiate text from huge amount of non-text elements.

SIFT. These features will not change with object position, orientation, scaling, illumination, noise/blur, and viewpoint.

SURF (Speed-Up Robust Features). It is another descriptor which is mostly used for features. This algorithm is grounded on theory of multi-scale space and feature detection. Feature detector is implemented by Hessian matrix. It performs well and display accurate results. Among the descriptors, SIFT is noble in performance.

Neural Networks [10, 57, 44, 59] have proved that feature learning which learns features from data is more effective in scene text. [10] Used SVM to classify features learned by unsupervised learning with a large filter size of 8x8 for the first layer. They have shown that more learning features could lead to better performance. [7] Have improved the system with convolution neural networks (CNNs) by using more training data and sharing features among case-sensitive, case-insensitive and bigram classifiers.

3.2 Tracking of Text in Video

In order to understand the visual content, spatio-temporal and multiple frame amalgamation analysis is immense strategies. Compared to imageries, the dependencies among adjacent frames of visual is beneficial in ameliorating discovery and recognition of text. In [3], spatio-temporal data obtained from various frames, tracking of text, tracking-based discovery and recognition approaches are expansively contemplated and emphasized.

The main task of tracing is to continually locate text across manifold dynamic visual frames. It is useful for verifying, integrating, enhancing text in video for detection and recognition process. [30] Proposed an evaluation strategy which improves gradually. They initially tracked the text by using the Sum of Square Difference (SSD) based matching. Apart from these, [3] have projected many other strategies like Mean Square Error (MSE), Least Mean Square Error (LMSE), and SAD (Sum of Absolute Difference) to diminish the complexity.

3.3 Text Extraction and Segmentation Techniques

Extraction of text is the process of segmenting components of text from background. Later, enhancement of text elements is done to avoid low-resolution and noise. The main intent of segmentation is, to segregate imagery to separate interest regions (video texts). Two approaches, namely, bottom-up approach and top-down approach segments each image into multiple separated regions. The former uses appropriate threshold to translate grayscale imagery to black & white imagery. But, selection of proper threshold value will be a difficult task. Histogram based methods are bottom-up method which computes a histogram by taking all the pixels in the image. Later, in the histogram, spotted peaks and valleys are used to find the groups in the imagery to perform segmentation. But the problem is from the histogram the single threshold value which is attained will lead to depletion of entity information in contrast to binarization of backgrounds. Below is the figure (Fig. 3) which lost the text after applying Otsu's method [45]:

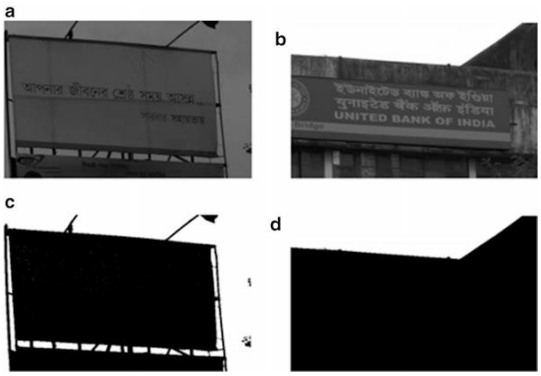


Fig. 3histogram based image binarization: above two is natural scene imagery, below two are the results of Otsu's method for binarization [61]

In [4], each text candidate is treated as cluster to excerpt full text line. Then, all clusters are grouped which represents text candidates. They have used Directional Region Grouping which discovers the direction of text clusters and assembles the nearest neighbour clusters. DRG works fine for any direction of text lines.

[5] Have used MSER algorithm to extract character candidates where most of the candidates which are not characters are minimized by pruning algorithm for MSERs.Non-text regions are abolished by linear reduction and tree accumulation algorithm using the procedure of minimal regularized variants.In [19], hybrid method is used, one is sliding window based method and another is CC (connected-component) together. Text region indicator is established to find the text predominant confidence and candidate region through binarization process.

[14] Proposed an algorithm which detects object with text like features like homogeneity of magnitude, color, stroke, breadth, space etc.; it gives false positives more. [15] Surveyed and concluded that discovery and localization of text process is not vigorous for discovering all classes of text. [16] Used natural language processing for recognizing text in videos where optical character recognition structure binarize the images first, and then extricate the visual features by classifier. Neural approach is used to learn for appropriate descriptors extraction and recognition of the character without binarization process. They have selected heterogeneous ConvNets configuration which consists of six hidden layers and an output layer. They have used linguistic knowledge to overcome the problem of character confusion, complexity of background, segmentation faults, and poor video quality for recognition.

[17] Proposed an algorithm which made hard to segment the text because of distortion in images, low resolution of videos. This led to poor recognition results. To overcome this, multi-scaled character symbol recognition and graphical model was proposed.

3.4 Text Recognition techniques

Text recognition is a process of identifying a text from scanned documents or images or video into a form which can be analysed and manipulated by a computer. OCR (Optical character Recognition) is the costly and speedy method for identifying the characters. Nowadays OCR engines include various neural network algorithms to study the strokes, gap between the characters from background. For each algorithm, the bright and dark along the side of a stroke gets averaged by allowing for irregularities of written ink on sheet, matches it to known characters. [19] Used OCR to verify the text after classifying text and non-text using a graphical model called CRF which depends on Markovian property.

Recent trends have transformed OCR to machine learning approach to identify the text from video. Machine learning allows creating architecture with multilayer perceptron to train the machine to categorize text and non-text. Once they are classified, machine learning techniques can be utilized for verification of text from non-text.

4. SPECIAL ISSUES

Although the papers which were reviewed till now achieved promising results, still detection of text in uncontrolled environments remains very challenging. Many have contributed methodologies towards the image-based text detection system.

Few papers were on scene text detection or natural scene detection. Relatively diminutive effort has been completed for extending these results to the visual field. For motion text, outcomes require well planned tracking procedures to assure precise text region at pixel level for registration. Analysis of Spatio-temporal for visual text discovery and/or recognition has been developed, too. Phuc focused on the issue of text detection in visuals. They have explored a diversity of approaches for training local character prototypes. They have also presented performance of detection on ICDAR 2013 visual dataset using Video Analysis and Content Extraction benchmark. They have also proposed a new metric for performance grounded on precision and recall curvatures to mensurate the performance of text recognition in visuals. They have shown the efficiency of exploiting inter-frame redundancy to eliminate false positives. This paper has extended the solution of [20, 21].

5. EVALUATION AND RESULTS

5.1 Dataset

[1] Surveyed performance of representative approaches. They have gathered frequently used datasets and recapitulated their attributes which includes text classes, sources, languages as well as training/testing samples. The MSRA, ICDAR 2011 and ICDAR 2013 datasets include graphics text in visual, web-based imageries and emails. The ICDAR 2003/05 and ICDAR 2011/13 datasets are prepared for natural scene text, text localization and character/word classification. ICDAR'13 dataset contains 28 visual sequences prepared to assess the visual scene text discovery, tracking and recognition.

ICDAR-VISUAL in the ICDAR'13 Robust ReadingChallenge 3 [51] presented a novel visual dataset to address the issue of text detection in visuals. This dataset contains 28 visuals, in which, 13 visuals are for training and 15 for testing. Each individual video contains captured scenes from live situations by using diverse cameras. Videos present a set of different challenges. In few videos, the quality of the imagery is worse than static imageries, due to blur, poor contrast, distortion, while visual compression might create further artefacts. The ICDAR 2013/15 datasets are prepared for text localization and end-to-end tasks. End-to-end tasks include localization of text followed by recognition of words.

YouTube Audio-visual Text YouTube Audio-visual Text (YVT) comprises of 30 videos. Each video has 15-second length, 30 fps, and HD quality 720p. Text in dataset could be distributed into dual groups, overlap text (captions, movie heading, and emblems) and natural scene text (motorway marks, commercial logos and symbols on shirt).

5.2 Performance metrics

Detection rate is a measure of performance for natural scene extraction. It is calculated as the number of detected texts divided by the number of texts for that frame. ICDAR 2003 shows precision/recall measures for retrieval system. Precision, p' is termed as ratio between the number of true estimates (TE) divided by the total number of estimates (E), p'= TE/E. System which shows over-estimations of the number of text rectangles will acquire low precision. Recall, r' is termed as the number of true estimates (TE) divided by the total number of targets (T), r'= TE/T. System which under-estimates the number of text rectangles will have a low recall.

Precision = number of true estimates / total number of estimates

Recall= number of true estimates / total number of targets

The ICDAR set has two drawbacks. First, most of the texts are horizontal. Next, all the texts exist in English.

Video Precision and Recall

Computing precision and recall for video is the challenging task compared to images. As video is a continuous motion picture, we need to track the video for accuracy in detecting the text. Tracking is a process of locating multiple texts over time. Much work has been done previously in estimating the accuracy of text detection and recognition from images but performing multi frame tracking accuracy with respect to text distortion, occlusions, blur, scale and rotation variant has become a challenging task.

Performance also depends on the number of parameters discussed by the technique to achieve the detection and recognition of text in video. Table I shows contemporary feature descriptors for spotting text in video with their advantages and disadvantages and parameters on which these are dependant for achieving tremendous performance.

6. DELIBERATION

Many traditional techniques have been explored from previous papers which have proven not efficient in detecting text from video. Also, many techniques were used to work with images and only few algorithms worked on videos. Later, computer vision, machine learning has taken tremendous lead to work for detection of video text with success rate. Mainly, MSER, SIFT and CNN took major role in detecting and recognizing the text. These three techniques are compared in the Table-I. This table also projects the dependant parameters for the above three mentioned techniques. The major idea behind our work is to alleviate the mentioned dependant parameters as much as possible, so that a novel algorithm can perform the detection and recognition process with more efficiency. SIFT and MSER use different methodologies utilize threshold values to find the stable key-points for feature detection. It is time consuming process to apply different threshold values and then analyse for stable regions. Also, as number of pixels of an image frame increases the complexity and accuracy of results will become cumbersome. These drawbacks can be overcome by either using fusion techniques followed by CNN or CNN with less numbers of parameters. Nowadays, Graphical Processing Unit (GPU) along with the CPU is utilizing to compute and accelerate deep learning process to enhance the performance of the system in analysing the voluminous data. In detection of text from video, deep learning process plays a vital role in processing speed, accuracy rate. Though, CNN uses GPU computing power, still it depends on many parameters like number of filters, receptive field and features which are

essential in modelling the system for analysis. These parameters take maximum time in pre-processing the video prior to feature extraction. Many research efforts have corroborated that fusing video text can progress video retrieval systems. Moreover, videos in web comprises of caption text and scene text, which are challenging because of their complex background and foreground attributes like distortion, illumination, orientation, skewness etc. Hence, the gigantic cumulative volume of videos in web requires real-time processing for textual recognition and retrieval.

Table I. Comparative Study of Ubiquitous Feature Descriptors in Text Spotting

I.a. Strength and Weakness

Descriptors	Strength	Weakness	Category
SIFT	Scale invariant More accuracy Rotation invariant High processing speed	High Computational cost Inefficient for low powered devices	Scene and caption text
CNN	Less formal statistical training Detect non-linear relationship between dependant and independent variables Detect all possible interactions	great computational cost Proneness to overfitting Need lot of training data Slow to train if GPU is not present	Scene and caption text
MSER+ sliding window-based method	Best for jpeg image Performs well for view point change Scale invariant	Not invariant with motion blur Cannot handle complex text information efficiently	Scene text

I.b. Dependant parameters of feature descriptors

Descriptors	Dependant Parameters	Description
SIFT	Number of pixels per key point Prior Smoothing Border distance Number of Difference of Gaussians Grayscale Threshold Ratio of curvature Radius of Gaussian Match ratio	Less number of key-point gives more accurate results Used to eliminate noise by blurring the original imagery at the beginning Distance between pixel(features) position and border, it detects the features when the distance is decreasing Used for key-points detection To discard low-contrast key-points Used to test if key-points are located on an edge. They are not reliable Orientation assignment Image alignment
MSER+ sliding window-based method	Delta value to detect maximally stable region Minimum area, maximum area Maximum variation Minimum diversity	stable regions are extracted which are maximally stable at different threshold stable region is rejected if it has less than minimum area pixels and more than maximum area pixels same as delta but smaller maximum variation gets less regions prune regions which are too similar
CNN	Input/output volumes Features Number of Filters Receptive Field Zero-padding	image constitutes a 3D structure in its entirety image analysis is done based on pattern obtained from input data operator applied on input image to transform the information encoded in the pixels pixels are fully connected to the NN's input layer size is adjusted to our requirement

7. CONCLUSION

This paper highlights the extensive survey on text detection from video. We have investigated numerous works and analysed techniques are exploited for text spotting. The procedure involved in text spotting can be divided into a) detection b) tracking c) segmentation d) extraction and e) recognition. There are many algorithms are available to detect and recognize the text from video. However, there is no unique method to satisfy scene text, caption text and both caption and scene text.

Although, many algorithms have referred and implemented to resolve the issue of detection and recognition of text, there is no efficient system yet for users. Earlier OCR was used to recognize the text. However, it is not appropriate to imageries and video. To apply in real time applications, we need to work with more efficient methods by make use of machine learning, hybrid techniques, fuzzy logic [23] etc. Also, this paper describes the datasets and performance

metrics used by different authors. This paper mainly conveys refined procedures to classify, analyse and review these algorithms, and discuss the future work to be done.

REFERENCES

1. Q. Y and D. Doermann, "Text detection and recognition in imagery: A survey," IEEE Trans. Pattern Anal. Mach. Intell., vol. 37, no. 7, pp. 1480–1500, Jul. 2015.
2. Y. Zhu, C. Y, and X. Bai, "Scene text detection and recognition: Recent advances and future trends," Frontiers Comput. Sci., vol. 10, no. 1, pp. 19–36, Feb. 2016.
3. Xu-Cheng Yin, Ze-Yu Zuo, Shu Tian, and Cheng-Lin Liu, "Text Detection, Tracking and Recognition in Video: A Comprehensive Survey," IEEE Trans.Image Processing, vol.25, no. 6, June.2016.
4. Liang Wu, PalaiahnakoteShivakumara, Tong Lu, and Chew Lim Tan, "A New Technique for Multi-Oriented Scene Text Line Detection and Tracking in Video," IEEE Trans.Multimedia, vol.17, no. 8, Aug-2015.
5. Xu-Cheng Yin, Xuwang Yin, Kaizu Huang, and Hong-Wei Hao, "Robust Text Detection in Natural Scene Images," IEEE Trans.2014.
6. Weilin Huang, Yu Q, and Xiaoou Tang, "Robust Scene Text Detection with Convolution Neural Network Induced MSER Trees," Springer, ECCV-2014, pp. 497-511.
7. Max Jaderberg, Andrea Vedaldi, and Andrew Zisserman, "Deep Features for Text Spotting," 2014.
8. Jing Zhang, and RangacharKasturi, "A Novel Text Detection System Based on Character and Link Energies," IEEE Trans. Image Processing.Vol.23, no. 9, Sep-2014.
9. Qixiang Ye, and David Daermann, "Text Detection and Recognition in Imagery: A Survey," IEEE Trans. Pattern Anal. Mach. Intell., vol.37, no.7, july-2015.
10. [10] Tao Wang, David J. Wu, Adam Coates, and Andrew Y. Ng, "End-to-End Text Recognition with Convolution Neural Networks," ICPR-2012.
11. Qixiang Ye, Qingming Huang, Wen Gao, Debin Zhao," Fast and robust text detection in images and video frames," Image and Vision Computing, Elsevier, Jan-2005.
12. Amit Panwar, HimanshuSuyal, "Fusion based text detection and localization for video image," IJRASET.Vol. 4 Issue.IV, ISSN: 2321-9653, April-2016.
13. Lukas Neumann, Jiri Matas, "Real-Time Scene Text Localization and Recognition," IEEE conference on CVPR-2012, June-2012.
14. S. Antani, D. Crandall, R. Kasturi, "Robust Extraction of Text in Video," ICPR-2000.
15. K.Jung, K.I.Kim, A.K.Jain, " Text information extraction in images and video: A Survey," Pattern Recognit., vol. 37, no. 5, pp. 977-997, May 2004.
16. K.Elagouni, C.Garcia, P.Sebillot, "A Comprehensive neural-based approach for text recognition in videos using natural language processing," in Proc. ACM Int. Conf. Multimedia Retr. (ICMR),2011.
17. K.Elagouni, C.Garcia, F.Mamalet, P.Sebillot, "Text Recognition in videos using a recurrent connectionist approach," in Artificial Neural Networks and Machine Learning- ICANN. Berlin, Germany: Springer, 2012.
18. Guozhu Liang, PalaiahnakoteShivakumara, Tong Lu, Chew Lim Tan," Multi-Spectral Fusion Based Approach for Arbitrarily Oriented Scene Text Detection in Video Images," in IEEE Trans. on Image Processing, Vol. 24, No.11, Nov-2015.
19. A.Thilagavathy, K.Aarthi, A.Chilambuchelvan, "Text Detection and Extraction from Videos using ANN based network," in IJSCAI, Vol. 1, No. 2, Aug-2012.
20. Phuc Xuan Nguyen, Kai Wang, Serge Belongie, "Video Text Detection and Recognition: Dataset and Benchmark," in IEEE winter conference on applications of computer vision (WACV), 2014.
21. K. Wang, B. Babenko, and S. Belongie," End-to-end scene text recognition," in ICCV, 2011.
22. K.Mikolajczyk and Cordelia Schmid, "A Performance Evaluation of Local Descriptors," IEEE Trans. Pattern Anal. Mach, Intell., vol. 27, no. 10, pp. 1615-1630, Oct 2005.
23. D Srinivasa Rao, M Seetha, MHM Prasad "Comparison of fuzzy and neuro fuzzy image fusion techniques and its applications" IJCA, vol 43, no.20, April 2012.
24. kopf J, "Capturing and viewing gigapixel images," ACM Trans Graph 26(3):93, 2007.
25. Hasan YMK, Karam LJ, "Morphological text extraction from images", IEEE Trans Image Process 9(11), pp. 1978-1983, 2000.
26. Jae-Chang S, Dorai C, B- R, "Automatic text extraction from video for content-based annotation and retrieval," Pattern Recogn., 1998.
27. Zhong Y, Karu K, Jain AK, "Locating text in complex color images," Pattern Recogn 28(10), pp. 1523-1535, 1995.
28. Ohya J, Shio A, Akamatsu S, "Recognizing characters in scene images," IEEE Trans, pp 214-224, 1994.
29. Kim HK, "Efficient automatic text location method and content-based indexing and structuring of video database," Commun Image Represent, pp 336-344, 1996.
30. Li H, Doerman D, Kia o, "Automatic text detection and tracking in digital videos," IEEE Trans, PAMI 9, pp 147-156, 2000.
31. Shivakumara P, Phan TQ, Tan CL, "New fourier-statistical features in RGB space for video text detection," IEEE Trans (TCSVT), pp 1520-1532, 2010.
32. Shivakumara P, Phan TQ, Tan CL, "A Laplacian approach to multi-oriented text detection in video," IEEE Trans (TPAMI), pp 412-419, 2011.
33. ViniVidyadharan, and SubuSurendran, "Automatic Image Registration using SIFT-NCC", Special Issue of International Journal of Computer Applications (0975 – 8887), pp.29-32, June 2012.
34. Luo Juan, and OubongGwun, "A Comparison of SIFT, PCA-SIFT and SURF", International Journal of Image Processing (IJIP), Vol. 3, Issue 4, pp. 143-152, 2009.
35. C. Yi, Y.L. Tian, "Textstring extraction from natural scenes by structure-based partition and grouping," IEEE Trans on Image processing 20 (9), 2011.
36. R. Kasturi, D.Goldgof, P.soundararajan, V.Manohar, J.Garofolo, R.Bowers, M.Boonstra, V.Korzhova, J.Zhang, "Framework for performance evaluation of face, text, and vehicle detection and tracking in video: Data, metrics, and protocol," PAMI, 2009.
37. B.Keni, S.Rainer, "Evaluating multiple object tracking performance: the clear not metrics," EURASIP JIVP, 2008.
38. A. K.Jain and B.Yu," Automatic text location in images and video frames," Pattern Recognit., vol. 31, no. 12, pp 2055-2076, 1998.
39. Wang.K and Kangas.J.A, "Character location in scene images from digital camera," Pattern Recognit., vol. 36, no. 10, pp. 2287-2299, 2003.
40. X. Chen. J.Yang, J.Zhang, A.Waibel, "Automatic detection and recognition of signs from natural scenes," IEEE Trans. Image Process., vol. 13, no. 1, pp. 87-99, 2004.
41. N. N and N.Papamarkos, "Color reduction for complex document images," Int. J. IMag. Syst. Technol., vol. 19, pp. 14-26, 2009.
42. A. Risnumawan, P. Shivakumara, C.S. Chan, C.L.Tan, "A robust arbitrary text detection system for natural scene images," Expert Syst. Appl., vol. 41, no. 18, pp. 8027-8048, 2014.
43. C. Yao, X.Bai, W.Liu, Y.Ma, Z.Tu, "Detecting texts of arbitrary orientations in natural images," CVPR, pp. 1083-1090, 2012.
44. X. C.Yin, W.Y.Pei, J.Zhang, Hong Wei Hao, "Multi-orientation scene text detection with adaptive clustering," IEEE Trans. Pattern Anal. Mach. Intell., vol. 37, no. 9, pp. 1930-1937, 2015.
45. W. Kim, C. Kim, "A new approach for overlay text detection and extraction from complex video scenes," IEEE Trans. Image Process., vol. 18, no. 2, pp. 401-411, 2009.
46. Yuqi Zhang, Wei Wang, Liang Wang, and Liuan Wang, "Scene Text Recognition with Deeper Convolution Neural Networks," IEEE, ICIP-2015.
47. Lienhart RW, Stuber F, "Automatic text recognition in digital videos," Proc SPIE, 2666(3), pp. 180-188, 1996.
48. Palma D, Ascenso J, Pereira F, "Automatic text extraction in digital video based on motion analysis," Image Analysis and Recognition, Springer, pp 588-596, 2004.
49. Box T, "High accuracy optical flow estimation based on a theory for wraping," ECCV, Springer, pp 25-36, 2004.
50. D. Karatzas, F.Shafait, S.Uchida, M.Iwamura, S.R. Mestre, J.Mas, D.F.Mota, J.A.Almazan, and L.P. de las Heras, " icdar 2013 robust reading competition," ICDAR, 2013.
51. R. minetto, N. Thome, M.Cord, J. Fabrizio, B. Marcotegui," A multiresolution system for text detection in complex detection in complex visual scenes," IEEE Conf, vol. 1, pp 3862-3864, 2010.

52. C. Garcia and X.Apostolidis, "Text detection and segmentation in complex color images," IEEE Int. Conf., pp. 2326-2330, 2000.
53. Lee.S, Cho.M, Jung.K, J.Kim, "Scene text extraction with edge constraint and text collinearity," IEEE Int. Conf.Comput. Vis. Pattern Recognit., pp. 3983-3986, 2010.
54. D. Ciresan, U.Meier, J.Schmidhuber, " Multi-column deep neural networks for image classification," IEEE Int. Conf. Computer Vision and Pattern Recognition, pp. 3642-3649, 2012.
55. M.Zeiler D, R.Fergus, "Visualizing and understanding convolutional netwotks," ECCV 2014, pp. 818-833, 2014.
56. A. Coates, B.Carpenter, C. case, S.Satheesh, B.Suresh, T.Wang, D.J.Wu, Andrew Y. Ng, "Text detection and character recognition of scene images with unsupervised feature learning," IEEE Intl Conf. Document Analysis and Recognition, pp. 440-445, 2011.
57. D. Kumar, M.N.A.Prasad, A.G.Ramakrishna, "Multi-script robust reading competition in ICDAR 2013," ICCV, pp. 569-576, 2013.
58. X. Rong, C.Yi, X.Yang, Y.Tian, "Scene text recognition in multiple frames based on text tracking," IEEE ICME, pp. 1-6, 2014.
59. L.Kang, Y.Li, D.Doerman, "Orientation robust text line detection in natural images," CVPR, pp. 4034-4041, 2014.
60. Szeliski R, "Computer Vision: algorithms and applications," Springer, 2010.
61. T. Lu, Shivakumara.P, Chew L.T, W. Liu, "Video text Detection," Springer, 2014. B
62. Evaluation metics for text extraction algorithms