IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Revue d'Intelligence Artificielle
Vol. 35, No. 5, October, 2021, pp. 383-393
Journal homepage: http://iieta.org/journals/ria

ETDR: An Exploratory View of Text Detection and Recognition in Images and Videos

Chaitra Yuvaraj Lokkondra[1*], Dinesh Ramegowda[2], Gopalakrishna Madigondanahalli Thimmaiah[3], Ajay Prakash Bassappa Vijaya[3], Manjula Hebbaka Shivananjappa[3]

[1] Department of CSE, Jain University, Bengaluru 560069, India
[2] Department of ISE, Jain University, Bengaluru 560069, India
[3] Department of CSE, SJBIT, Affiliated to Vivesvaraya Technological University, Bengaluru 560060, India

Corresponding Author Email: ylchaitra@gmail.com

https://doi.org/10.18280/ria.350504

Received: 27 September 2021
Accepted: 15 October 2021

Keywords:
text detection, text recognition, machine learning, deep learning, benchmark datasets

ABSTRACT

Images and videos with text content are a direct source of information. Today, there is a high need for image and video data that can be intelligently analyzed. A growing number of researchers are focusing on text identification, making it a hot issue in machine vision research. Since this opens the way, several real-time-based applications such as text detection, localization, and tracking have become more prevalent in text analysis systems. To find out more about how text information may be extracted, have a look at our survey. This study presents a trustworthy dataset for text identification in images and videos at first. The second part of the article details the numerous text formats, both in images and video. Third, the process flow for extracting information from the text and the existing machine learning and deep learning techniques used to train the model was described. Fourth, explain assessment measures that are used to validate the model. Finally, it integrates the uses and difficulties of text extraction across a wide range of fields. Difficulties focus on the most frequent challenges faced in the actual world, such as capturing techniques, lightning, and environmental conditions. Images and videos have evolved into valuable sources of data. The text inside the images and video provides a massive quantity of facts and statistics. However, such data is not easy to access. This exploratory view provides easier and more accurate mathematical modeling and evaluation techniques to retrieve the text in image and video into an accessible form.

1. INTRODUCTION

As the number of available multimedia documents grows, so does the demand for information classification and retrieval. As a result, text retrieval in images and videos has become more popular. Applications such as mobile text detection, sign recognition and transformation, license plate interpretation, and content-based image finding are just a few examples. Text is made up of parallel specified models of concepts expressed in language. Text objects integrated into video offer a wealth of helpful information about the multimedia system. Text extraction techniques are crucial in the classification and retrieval of multimodal knowledge. The applications, such as record processing, image retrieval, video extraction extracting text from images or videos, can be a significant disadvantage [1, 2]. Typically, words embedded in an extremely image or a framed record crucial media setting such as the player's details, film titles, news channel dates, article introduction, and so on.

The rapid use of Smartphone's and online social media has resulted in gathering an enormous volume of pictorial data, particularly the enormous and growing gatherings of videos on websites and social media. There is a significant need for effective indexing and critical information retrieval from videos and the segmentation of text area. Several techniques were explicitly created to make this work simple [3]. In 2014, YouTube streamed almost 100 hours of video in every minute throughout the world. These videos have sparked study in multimedia comprehension and video examination [4, 5]. Text is one of the most expressive forms of communication. It may be used to communicate information in papers or situations. This is done in such a way that it will be "noticed" and readable by others. The demand for increasingly numerous applications in various fields is the primary cause of this development. Some other reasons are highlighted in the following content: first, gathering vast volumes of "street view" data. Second, high-performance mobile devices with image and processing capabilities are becoming more widely available [6, 7]. Due to the ease of image capture, text recognition is possible in a variety of settings. Third, the progress in the field of machine vision technology makes it easier to solve complex issues.

Many of the challenges for word recognition and text identification in images are similar to those of machine vision and pattern detection issues; these are exacerbated by low-quality or degraded data. Current methods have less detection level (sometimes < 80%) and recognition level (typically < 60%) [8, 9]. A scannable document's OCR recognition rate, on the other hand, is generally greater than 99 percent [10]. The background complexity and differences in text arrangement, poor resolution, multilingual material, and uneven lighting present a bigger problem than clean, well-formatted documents. The deployment of sophisticated computer vision and pattern recognition algorithms is required to solve these challenges. Identifying and recognizing text in the image has been the topic of several research proposals over the past five years. However, we are unaware of any extensive studies on the issue. However, the majority of the examined material was

published before 2003.

Video-text has become a more important source of information [11]. Several other ways to utilize text as a visual signal for navigation and notification include caption-text [12] and signpost text [13]. A great deal of attention is now being paid to text extraction and analysis in video. For example, when it comes to video retrieval, all recent winners of the Multimedia Event Detection TRECVID2 competition have used a combination of text, audio, and visual characteristics. The capacity to interpret video text can significantly increase retrieval performance. According to some studies, video retrieval utilizes textual (extracted from frames and audio) and visual representations that employ high-level object and action ideas [14, 15].

Many surveys have been done so far in the text detection and recognition field. Most of the review papers take anyone field like extract text from image or video. And some other detailed the available techniques in either machine learning or deep learning. However, this paper gives the complete package of how to extract text from images and videos. The paper is detailed from scratch, from data collection to the final goal evaluation methods. This domain addressed two advanced techniques like machine learning and deep learning methods used by researchers so far.

The following is how the survey is put together: The dataset for image and video processing is detailed in section 2. The 3rd section categorizes the many forms of text in images and videos. The 4th section looks into the step-by-step method of extracting data from images and videos. In section 5, the validation procedures are described in depth. Sections 6 and 7 discuss the application and difficulties.

2. DATASETS

For developing a great model for better results, good quality data and textual annotations are necessary. The most popular benchmark datasets of images are shown below. Table 1 offers a complete list of all of them.

Table 1. Details of benchmark datasets for text detection and recognition from images

Datasets	Year	Language	Multi-oriented	Arbitrary shape	Total Samples
IC03	2003	English	NO	NO	484
SVT	2010	English	NO	YES	350
KAIST	2011	Multi-lingual	NO	NO	3000
M500	2012	English/ Chinese	NO	YES	500
IIIT 5K	2012	English	NO	YES	5000
IC13	2013	English	NO	NO	462
IC15	2015	English	NO	YES	1500
COCO-Text	2016	English	NO	YES	63686
SynthText	2016	English	NO	NO	858750
CTW	2017	English/ Chinese	NO	YES	1500
MLT17	2017	Multi-lingual	NO	YES	18000
RCTW-17	2017	English/ Chinese	NO	YES	12514
ToT	2017	English	YES	YES	1525
LSVT19	2019	English/Chinese	YES	YES	450000
MLT19	2019	Multi-lingual	NO	YES	20000
ArTs19	2019	English/Chinese	YES	YES	10166

The ICDAR2003 [16] is a scene text recognition dataset. It contains 258 training examples and 249 testing examples of natural scenes. In this dataset, Character-level annotations are used for the images. Images can be altered to extract important text and characters. Street View Text (SVT) [17, 18] comprises 350 images from Google Street View annotated with bounding boxes in this set of 350 images. In addition, not all text occurrences are annotated. In this dataset, 17,548 text occurrences are included. KAIST [19] contains three-hundred-and-fifty images that were shot in a variety of environments and lighting situations. For taking real-world images, a well-resolution camera or a Smartphone camera is utilized. Every single image has been scaled to the standard resolution of 640 x 480. M500 (MSRA-TD500) [20] dataset comprises 500 different nature scenes, with 300 images used to train and 200 images used to test. Additionally, it allows annotating text at the line level and using polygon boxes to do so. Both English and Chinese texts may be found within.

The IIIT 5K [21] data collection is derived from Simple Google searches. The film poster, billboard, sign plate, nameplate, numbers in the address display were used to gather images. Over 5000 words have been clipped from Scene Texts and digital images present in this dataset. The collection is split into two sections: training and testing. This set of data may be used to recognize words from a vast vocabulary. In addition to the dataset, they also supply a lexicon of almost 0.5 million vocabulary terms. The ICDAR 2013 (IC13) [22] trains on 229 nature images and tests on 233. ICDAR 2015 (IC15) [23] holds 1,500 images, 1,000 are used for training, and 500 are used for validating or testing. Four quadrilateral vertices are used to annotate the text, which is generally distorted or obscured because the material was collected without the user's consent or knowledge.

COCO-Text [24] is the significant dataset for text detection and recognition developed up to date. The 43,686 images were employed for learning and 20,000 for evaluating, which includes handwritten or printed text and clear and fuzzy text in English or non-English. SynthText [25] Eighty-eight thousand seven hundred and fifty synthetic images have been created to provide a realistic look. At the character and word levels of annotation, the text in this dataset is broken down into lines and words. SCUT-CTW1500 [26] holds 10,751 cropped word images in this dataset with 1,500 images, including 1,000 for learning and 500 for evaluating. A polygon with 14 vertices is what CTW-1500 uses for its annotations. Words in Chinese and English make up the majority of the dataset. In the multilingual text of ICDAR 2017 (MLT17) [27], there are 20,000 natural scenery images in the dataset, including 7,200 data for training, 1,800 data for validation, and 9,000 data for testing. It gives word-by-word commentary. RCTW-17 [28] dataset is used for events on identifying and finding the English/Chinese language in images; this dataset comprises a variety of images, including street points, banners, inside settings, and screenshots of the images. Annotations are identical to those in ICDAR2015. Compared to prior datasets, the Total-Text (ToT) [29] has a higher proportion of curved text than previous ones. Most of these images were sourced from street boards and labeled as geometric shapes with varying vertices. A large-scale street view text (LSVT) of ICDAR 2019 [30] provides annotations to all 20,000 test images, 30,000 training images. ICDAR 2019 (MLT19) [31], this dataset comprises 18,000 images that have been annotated at the word level. There are ten languages in this dataset. For scene text detection, it is a more simple and complicated set of data. ArTs19 (ICDAR 2019 Arbitrary-Shaped Text) [32]

comprise 10,166 images, 5603 images are used for training and 4,563 for testing. Text shapes of various kinds (horizontal, multi-oriented, curved) are represented by many instances.

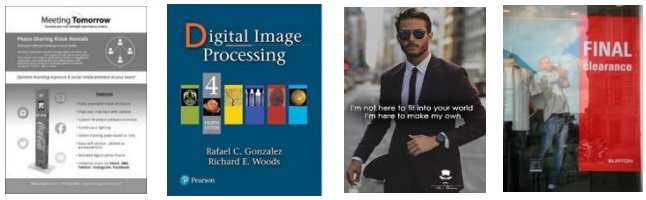
Table 2. Details of benchmark datasets for text detection and recognition from videos

Datasets	Year	Category	Source	Task
TREC	2002	Scene Text / Caption	Video Frames	Segmentation
SVT	2010	Scene Text	Video Frames	Detection and Recognition
MSR-VTT	2011	Scene Text	Video	Detection, Tracking, and Recognition
ICDAR2013	2013	Scene Text	Video	Detection, Tracking and Recognition
Merino-Gracia	2014	Scene Text	Video	Tracking
YouTube Video	2014	Scene Text / Caption	Video	Detection, Tracking and Recognition
ICDAR 2015	2015	Scene Text	Video	Detection, Tracking and Recognition
RoadText-1K	2020	Scene Text	Video	Detection and Recognition

The most popular benchmark datasets of videos are shown above. Table 2 offers a complete list of all of the video datasets in English languages. Graphic text in most videos travels at fast rates in this dataset. The text in specific videos is impacted by noise, distortion, blurring, and significant variations in lighting or occlusion. Aside from that, there are several other video frame databases-embedded text or scene text in the TREC [33] and SVT [18] datasets. For video text search, the TREC dataset is utilized. MSR-VTT [34] contains almost 10,000 data with 20 twenty categories. All videos are annotated in the English language. The data is divided as 6,513 and 497 from the whole dataset. The 28 videos in the dataset from ICDAR 2013 [22] are used to practice the detection, tracking, and identification of video scene text. Thirteen of the videos in this collection are for training, while the others are for testing. There are various scripts and languages represented in these films (including English, Spanish, Japanese and French) and a variety of camera types. Merino-Gracia [35] dataset has been developed to monitor scene text in the video. Scenes with text are included in the Merino-Gracia dataset. These scenes were shot with unpredictable handheld motion and suffered from a lot of blur and perspective distortions. There are considerable perspective shifts and viewpoint changes in the Merino dataset as well. SVT dataset is utilized to identify and recognize street view text [18]. Using the YouTube Video Text databases, it is also possible to identify, track, and recognize text in the video [36]. The YouTube Video Text dataset, in particular, consists of 30 videos gathered from YouTube. The title song, captions, and logos can be further separated. Recently, the ICDAR 2015 [23] unconfined a modification of the ICDAR 2013 data set. There are 25 videos (Total 13450 frames) in the ICDAR 2015 training dataset and 24 videos (Total 14374 frames) in the ICDAR 2015 test set. Organizers from several nations collected the data, which contains information in a variety of languages. In both indoor and outdoor settings, the video sequences correlate to seven high-level activities. This is done by using four separate cameras for various sequences. The RoadText-1K contains 1000 driving videos with text. This dataset is approximately twenty times larger than the previously available dataset for text detecting videos [37].

3. TEXT IN IMAGES AND VIDEOS

Page layout analysis worked on documents. The text exists in images, and videos are not in the same format. The various kinds of text available in images are represented in Figure 1 [38]. Figure 1a represents the text in a grayscale document image [36]. Figure 1b represents the text in a multi-color document image. This type of text differs from Figure 1 by color only. Then the Figure 1c is the image with a caption. In this type, the text is artificially added to the image. Finally, Figure 1d represents the scene text. Here the text is naturally present in the image.



(a) (b) (c) (d)

Figure 1. Types of text in images: (a) Grayscale document images, (b) Multi-color document images: (c) Images with the caption, and (d) Scene text images

A videos description text is divided into a scene text and caption text as shown in Figure 2 [36, 38]. Text used in captions is often denoted as fake or graphic text. This caption is again divided into two types: layered type and embedded type as shown in Figures 2a and 2b. Text in the caption or subtitle of the video gives clear direction and an overview of the information. In contrast, the text in scenes [39] illustrated in Figure 2c is captured by the camera; here, the text is organically integrated inside objects or scenes. For better understanding, the layered caption text is mostly on top of an appropriately designed background layer. In contrast, the embedded caption is embedded in the image. There is a consensus that scene texts and embedded captions are more challenging to discern.

(a) (b) (c)

Figure 2. Types of text in videos: (a) Caption type layered, (b) Caption type embedded, and (c) Scene type

4. METHODOLOGY

Recently, several approaches for interpreting text from images have been presented. For example, these approaches locate text regions in images by estimating bound boxes for all conceivable text areas and then identifying all discovered regions' content. Text detection and identification are two different jobs; reading text from images is broken down into two distinct steps. As the name implies, text detection is the process of identifying or localizing text in images. Contrarily, text recognition focuses exclusively on converting recognized text areas into editable letters, words, or text-line. The flow diagram and available techniques for extracting text

information from the image using machine learning and deep learning are shown in Figures 3 and 4.

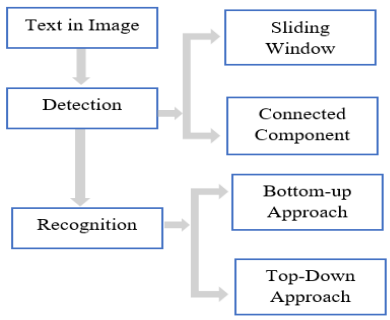


Figure 3. Flow diagram for extract text information from the image using machine learning

A. Text detection using machine learning

Scene text detection methods are summarized in this section and maybe parted into primary categories such as sliding window and component-based methods. As an example, Kim et al. [40] used a given test image to build a pyramid of images that can scan all potential text positions and scale them by utilizing a sliding window of a specific shape to scan over all possible text sites. A traditional classifier (AdaBoost and random forest) detects text in each window to classify certain image information and extract image regions with similar properties (such as color, texture, boundary, and corner points). Connected-component-based methods create candidate components that will be classified into text or non-text classes [41] using a conventional classifier like Random Forest, Support Vector Machine, and nearest-neighbor. These approaches extract characters from an image and connect them into a word, depending on the method. They are more efficient and resilient than sliding-window techniques, and they tend to have a reduced false-positive rate [FPR], critical in-text identification. Both MSER (Maximally stable extremal regions) [42] and SWT (Stroke Width Transform) are connected-component-based techniques that form the basis of many later text detection efforts.

B. Text recognition using machine learning

For example, identifying specific characters or words in a scenario may transform them into characters or words. Categorizing characters use case-sensitive characters, such as ten numbers, lowercase, uppercase letters of 26 each, 32 ASCII. A statistical language model is then employed to prune out misclassified characters in traditional scene text recognition methods that have been around for two decades. These methods combine typical image features, such as Scale-Invariant Feature Transform (SIFT) and Histogram of Oriented Gradients (HOG), with machine learning, like K-Nearest Neighbors and Support Vector Machine. A natural representation of a bottom-up approach offers a collection of important image areas individually described by the feature vector. The top-down approach predicts an interest allocation over image areas based on task-specific information. After that, the attentive feature vector is calculated as a weighted averaging of all image attributes across all areas. A bottom-up technique is used in most machine learning-based approaches, where categorized characters are combined as words. So, HOG features are retrieved from the sliding window, followed by a machine learning algorithm that has been pre-trained to identify the input word image's characters [43]. In other words, the word is identified directly from the input images rather than detecting and identifying individual letters [44].

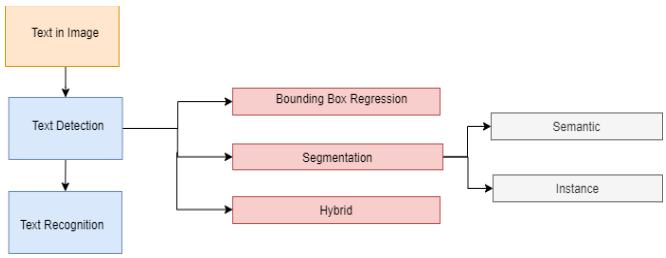


Figure 4. Flow diagram for extract text information from an image using deep learning

C. Text detection using deep learning

The effectiveness of past deep learning approaches may suffer when text is located inside a complex background, such as when features of the background have similar appearances to text. As shown in Figure 4, now deep learning-based text recognition techniques are separated into three groups: bounding-box regression, segmentation, and hybrid approaches to overcome the above problems.

Text identification approaches based on bounding-box algorithms [45] treat text as an object and effectively anticipate bounding boxes. The works [46] suggested a Fully Convolutional Network (FCN) based on the YOLO network [47] and also a random-forest technique to decrease inaccuracy levels in predicted images. To identify arbitrary-oriented text, Ma et al. [48] proposed Rotation Region Proposal Networks, which are based on Faster Region-Based Convolutional Neural Networks (R-CNN) [49]. TextBoxes++ was created by extending TextBoxes and enhancing the network model and training procedure. To identify arbitrary-oriented text, Textboxes++ changed the rectangular bounded box to a quadrilateral bounded box around text [50]. The segmentation method [51-53] intends to classify text areas in images as pixels. Such approaches retrieve textual information from a segmentation map created by an FCN and then construct bounding boxes for the text. The FCN was used to forecast a salient map of text areas and the center of each word in an image. The updated FCN generates three score patterns from the input images: text/non-text areas, character classes, and orientations. The bounding boxes with the segmentation maps are then subjected to a word separation post-processing method. These segmentation-based approaches work well with flipped and unstructured text; they may not be able to distinguish between adjacent-word occurrences. Several recent papers [54] have handled text identification as a segmentation problem. Many research studies employ the Mask R-CNN [55] structure to strengthen text identification performance, which helps detect text occurrences of various shapes. Hybrid techniques [56, 57] combine a segmentation-based methodology for predicting text score maps with regression to produce text bounding boxes. The pyramid mask text detector (PMTD) is a novel Mask R-CNN-based framework that allocates a soft pyramid label for all pixels in a text [58]. Differential Binarization Network (DBNet) was proposed for scene text recognition in the real world [59]. The

usage of an adaptive Bezier curve network is presented as a method for scene text detection [60].

D. Text recognition using deep learning

Many research suggested deep learning-based methods to solve the difficulties of recognizing text in the image. The study [61] suggested a CNN-based framework for extraction features to recognize the character in the text and then used the Non-maximum Suppression approach [62] to estimate the overall word. They developed an automated technique of binarizing color text areas in videos based on a particular CNN. An FCN was employed to encode character features, and subsequently, an n-gram method was used to recognize characters [63]. The research publication [64] used CNN architecture to perform a 90 thousand English word categorization. However, this approach outperformed for word recognition; it has some disadvantages; one is Out-of-vocabulary words and distortions of the lengthy word are not recognized. The ResNet model was used after extracting features from a CNN to forecast feature patterns [65]. A novel R-CNN-based architecture, Gated RCNN (GRCNN), is presented [66], utilizing a gate to regulate recurrent connections in an RCNN model. For dealing with text abnormalities, Liu et al. [67] suggested a spatial-attention residue Network (STAR-N) [68]. For identifying corrupted characters from the Character-Aware Neural Network (Char-Net) is presented [69]. The focus attention convolution LSTM (FACLSTM) for text recognition [70] was suggested after considering scene text recognition as a spatial-temporal prediction issue. A dynamic log-polar transformer and a sequence recognition network are combined to build a novel scale-adaptive orientation attention network for recognizing the randomly aligned text in an image [71]. This is useful for generating rotation and scaling awareness in visual representations.

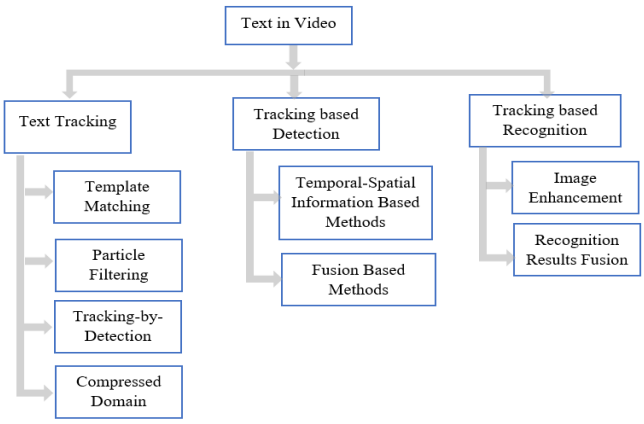


Figure 5. Flow diagram for extract text information from video

Multi-frame integration and spatial-temporal analysis are two of the most common methods used in video content comprehension; text detection and identification can be improved by using temporary information instead of images. Accordingly, video text tracking, tracking-related detection, and tracking-related identification techniques are discussed in this part, employing information from several frames. The flow diagram and available techniques for extracting text information from the video [11] are shown in Figure 5.

E. Text Tracking

With text tracking, the objective is to constantly detect text placement over many dynamic video frames in real-time. It is beneficial for verification, integration, enhancements, and speedup for text detection and identification. Different text tracking approaches have been studied in the literature in recent years. Most of these text monitoring approaches may be grouped into one of two categories: detection-based tracking and recognition-based tracking. A small study has been done on this subject to track the text and aggregates the text recognition results from many frames. The research [72] identifies the distance for a detected character in the current frame and a suggested character in the subsequent frame as a text-matching feature. This type of tracking relies on using objects or sensed positional data to track text over successive frames. According to standard tracking methodologies [73], text tracking using template matching, particle filtering, and detection tracking may be further divided into numerous subcategories.

F. Tracking Based Detection

Too far, the most common approach of finding videotext has been to look for text inside each frame or a few critical frames. A complicated background, low contrast, and lossy text compression prevent these techniques from achieving high detection accuracy. It is important to note that a high temporal repetition characterizes video to text. To minimize the number of false alarms and to increase detection accuracy, text tracking methods are used. These tactics are referred to as tracking-based detection strategies. Generalized temporal-spatial information [74] and fusion techniques [75] can classify these methods. In the first technique, the noise is removed directly by using temporal or spatial information. It is also possible to integrate detection and tracking findings in a single frame to increase accuracy. Background areas can help video text detection, according to the journal [76]. It only considers short-term dependencies. There are specific structures, such as optical flow [77], Conv3D [78], and Conv-LSTM [79], that are effective at capturing spatial-temporal information in object tracking. As a result, the article [80] incorporates a memory function within the architecture. This study shows that an end-to-end text detector in video performs better with the aid of long-term spatial-temporal memory and live tracking.

G. Tracking Based Recognition

Even though numerous text recognition algorithms have been developed, they have always been confined to single-image recognition. For a long time, recognition technologies such as Optical Character Recognition methods and additional products have been widely utilized to detect images' text. However, they cannot attain excellent outcomes in every video frame due to poor contrast, poor resolution, and complex backgrounds. Image enhancement [81] uses techniques like averaging multi-frame, finding minimum and maximum pixel intensity. The recognition results fusion [82] usually combines recognized text output of various frames into single final text output to produce a high-resolution image. If the region of text has been tracked, advanced tracking methodologies are employed. These two methods are generally used for proper recognition: (1) selection methodology by picking the correct

text occurrences from tracing regions and (2) fusion methodology by integrating successive recognition outcomes. Various recognition outcomes for the same text region are combined to get an end outcome, depending on performance metrics [83]. To separate a textual surface from the complicated background, the author offers a unique FCM-based extraction technique to retrieve text information. To represent every pixel for clustering, create a 5-dimensional feature vector that comprises location and color details. The text retrieved using the suggested approach is better and much more accurate when compared to previous techniques (Otsu and K-means). In article [84], a combo of CNN and LSTM has been used to detect Arabic textual data in a video sequence. Compared the efficiency of various pre-trained ConvNets to recognize and detect the text with the deep learning-based approach [85].

5. EVALUATION METRICS

The Precision and Recall techniques are employed to validate the performance of text detection models [22, 23]. The H-mean is used, and it is given in Eq. (1). The G in the following equations represents the ground-truth text, and D represents the detection text.

$$H_{mean} = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{1}$$

where calculating the precision and recall using,

$$Recall(G, D) = \frac{\sum_{i=1}^{|G|} PerfectMatch_G(G_i)}{|G|} \tag{2}$$

$$Precsiom(G, D) = \frac{\sum_{j=1}^{|G|} PerfectMatch_D(D_i)}{|D|} \tag{3}$$

where,

$$PerfectMatch_G(G_i) = max_{i=1\ldots|D|} \frac{2Area(G_i \cap D_j)}{Area(G_i) + Area(D_j)} \tag{4}$$

$$PerfectMatch_D(D_i) = max_{i=1\ldots|G|} \frac{2Area(D_j \cap G_i)}{Area(D_j) + Area(G_i)} \tag{5}$$

Word recognition accuracy (WRA): It is used in our everyday lives instead of character recognition accuracy. Word recognition accuracy (WRA) is a common statistic for evaluating text recognition methods [86, 87]. WRA is defined as follows given a collection of clipped word images:

$$WRA = \frac{Total\ correctly\ predicted\ words}{Total\ words} * 100 \tag{6}$$

The same evaluation metrics are used for video detection and recognition. For tracking, the different evaluation metrics are detailed [88].

MOTP is estimated by the spatial-temporal coincide between the process output and the reference trajectory.

$$MOTP = \frac{\sum_{i=1}^{N_{mapped}} \sum_{t=1}^{N_{frames}^{(t)}} \left[\frac{|G_i^{(t)} \cap D_i^{(t)}|}{|G_i^{(t)} \cup D_i^{(t)}|}\right]}{\sum_{t=1}^{N_{frames}} N_{mapped}^{(t)}} \tag{7}$$

where,
$G_i^{(t)}$: i[th] ground truth text at t-frame;
G_i: i[th] sequence ground truth text;
$D_i^{(t)}$: i[th] tracked text at t-frame;
D_i: i[th] sequence tracked text;
N_{frames}: Total frame;
$N_{frames}^{(t)}$: Total frame where ground-truth or tracked text exist in t-frame;
N_{mapped}: Total ground-truth and tracked text;
$N_{mapped}^{(t)}$: number of mapped ground-truth and tracked text pairs in frame t.

MOTA can be identified as the ratio of total false negatives and positives and switches to the ground-truth words at t-frame.

$$MOTA = 1 - \frac{\sum_{t=1}^{N_{frames}} (fn_t + fp_r + id\_swt)}{\sum_{t=1}^{N_{frames}} N_G^{(t)}} \tag{8}$$

where,
fn_t: Total false negatives;
f: Total false positives;
id_swt: Switches;
$N_G^{(t)}$: Ground-truth words at t-frame.

The Sequence Track Detection Accuracy (STDA) is a tracking performance measure for all the text in a particular order and is estimated by using the given formula.

$$STDA = \sum_{i=1}^{N_{mapped}} \frac{\sum_{t=1}^{N_{mapped}} \left[\frac{|G_i^{(t)} \cap D_i^{(t)}|}{|G_i^{(t)} \cup D_i^{(t)}|}\right]}{N(G_i \cup D_i \neq \emptyset)} \tag{9}$$

The normalization of Sequence Track Detection Accuracy per text is declared as ATA, the formula for computing the ATA is given as:

$$ATA = \frac{STDA}{\left[\frac{N_G + N_D}{2}\right]} \tag{10}$$

6. APPLICATIONS

Numerous text-based applications for images and video would be developed for the past 20 years and maybe roughly classified as industrial automation, multimedia retrieval.

A. *Retrieval of multimedia information*: Text in images on online pages is related to the content of the web pages [89, 90]. Typically, video subtitles include information on where, when, and whom the activities are taking place. In such multimedia resources, text recognition and keyword extraction improve multimedia retrieval. As mobile Smartphone's with digital cameras have proliferated, imaging equipment has become more readily available. This module automatically allows mobile devices to enter name cards, slides, and other

presentations [91-94]. Users are more comfortable and productive when they are not compelled to use a keyboard.

B. *Industry Automation*: There are several uses for text recognition in industrial automation. Mail sorting systems, for example, recognize addresses on envelopes. Automated tracking of container numbers helps logistics. Automatic geocoding systems gain from the ability to recognize house numbers and text on maps. There is much information that can be gleaned from signs in nature. Users can overcome language obstacles with the help of automatic sign recognition and translation software [63].

As mobiles, intelligent wearable gadgets, and social media have grown in the recent decade, video text retrieval techniques and systems have found a wide range of applications. Video understanding and retrieval and reading nature are two of the primary types of these applications that we will briefly discuss in this paper.

1. *Video Understanding and Retrieval*: As part of semantic-based video analysis and indexing, text in video plays a crucial role in retrieval and indexing [95]. Numerous studies have been done on this subject. To assist individuals and verify if their customers' ads have been shown on the TV channel in a given period, text identification and recognition were used to store the airing date and time of commercials. Video comprehension and retrieval will assist us in the future in acquiring more accurate and richer video content.

2. *Reading the nature*: With the rapid growth of text extraction techniques [96], numerous real-world claims such as aiding visually challenged individuals and identifying text in the roads have emerged and been developed. We have real-time translation, medical service, navigation system for the end-user, traffic monitoring, and assistive driving systems.

7. CHALLENGES

Text is usually set out in a way that makes it easier to read. According to the following categories and analyses, the problems posed by complex settings, a variety of image capture techniques, and the variability of text content are numerous.

Uneven lighting: Because of the illumination and the unequal reaction of sensors, lousy lighting is typical while collecting images. Color distortion and degradation of visual characteristics are introduced by uneven illumination, resulting in erroneous detection, recognition, and tracking outcomes.

Blurring and degradation: Defocus and clouding of text images occur when working conditions are flexible and not focused on cameras [92]. Text, particularly graphical video text, is degraded by image/video compression and decompression techniques. Defocusing, blurring, and degradation have the effect of reducing character sharpness and introducing touching characters, making fundamental operations like segmentation harder.

Distortion: If the ophthalmic axis of the lens is not vertical to the text plane, perspective distortion occurs. Rectangular forms disappear from text borders, whereas character shapes change, reducing the quality of identification models trained using undistorted data.

Aspect ratios: Text on traffic signs, text on video subtitles, and the size will vary. The text has multiple aspect ratios, in other words. A search strategy must be evaluated in terms of position, scale, and length to identifying text, which adds to the computational complexity.

Fonts: As a result of character overlap in italic typefaces, segmentation is complex. A vast number of character classes in different typefaces have many variances within each class and many pattern sub-spaces, which makes it challenging to recognize characters with accuracy.

There are several similarities between detecting and identifying text in the video, like resilience to background complication, blurring and misalignment of the text, and text distinction and mobile objects. Text detection and identification in images and videos have been widely explored in the literature [97]. However, this review focuses on many major video-based problems for text retrieval. These problems may be broken down into three primary categories: background-based, foreground-based, and video-based difficulties. Most of these video-based problems may be traced back to the following specific properties of dynamic videos:

- Compressed degradation: There is a risk of text color leaking when storing videos with lossy compression, which is typical. It is also possible to lose contrast between text and background when using low-bit-rate video compression. Aside from that, video compression may introduce new artifacts into the image.
- Low resolution: Video frames have a lesser resolution than paper images.
- Real-time processing: Most video clips have more than 25 and less than 30 frames per second. So, video text identification and tracking techniques require a lot of computing power. Most existing approaches and systems in the literature are unable to perform at such rates.
- Moving text and objects: For example, when the camera is zooming and rotating, or the scene text is moving randomly, the text might appear to be moving nonlinearly in a complicated fashion. There is also a tendency for moving objects in video to seem blurry.

8. CONCLUSION

Text information detection and recognition from image and video is the current trending research domain because of its fast growth applications in most fields. For doing research, the data is essential; this article gives reliable and accessible websites for collecting images and videos with text information. The text integrated or printed in an image of various forms is referred to as Image Text. Recorded photos, digitized articles, journals, newspapers, and advertisements all include visual text. Such texts are commonly accessible presently and thus play a significant role in expressing, explaining, and transmitting information to guide humans in interaction, issue resolution, creating jobs, cost efficiency, profitability, globalization, and cultural gaps, among other things. License plate recognition, self-navigating automobiles, vehicle tracking, industrial automation, and other applications need text extraction from videos. Text Extraction is a method of converting image or video text into ordinary text. Text Extraction helps extract information from image text for exploring, altering, recording, preserving, and evaluating it. The text extraction is a difficult task because of variations in dimension, alignment type, and font, and also text contained in diverse multicolored document images, damaged documents pictures, poor quality photos, poor visual brightness, and background clutter. This survey offers a

complete overview of image text detection, recognition, and tracking—the process flow to identify the model for extracting the information from raw images and video in detail. Then suggest the available evaluation metrics to measure the model quality. The applications and practical problems are also described in this study.

REFERENCES

[1] Zhang, X., Sun, F.C., Gu, L. (2010). A combined algorithm for video text extraction. 2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery, pp. 2294-2298. https://doi.org/10.1109/FSKD.2010.5569311

[2] Gopalan, C., Manjula, D. (2008). Contourlet based approach for text identification and extraction from heterogeneous textual images. International Journal of Computer Science and Engineering, 2(4): 202-211.

[3] Hu, W., Xie, N., Li, L., Zeng, X., Maybank, S. (2011). A survey on visual content-based video indexing and retrieval. IEEE Trans. Syst., Man, Cybern. C, Appl. Rev., 41(6): 797-819. https://doi.org/10.1109/TSMCC.2011.2109710

[4] Sharma, H., Kumar, S. (2016). A survey on decision tree algorithms of classification in data mining. International Journal of Science and Research (IJSR), 5(4): 2094-2097.

[5] Wang, L.S. (2019). Research and implementation of machine learning classifier based on KNN. IOP Conference Series: Materials Science and Engineering, 677(5): 052038. https://doi.org/10.1088/1757-899X/677/5/052038

[6] Chen, X.L., Yang, J., Zhang, J., Waibel, A. (2004). Automatic detection and recognition of signs from natural scenes. IEEE Transactions on Image Processing, 13(1): 87-99. https://doi.org/10.1109/TIP.2003.819223

[7] Liu, X. (2008). A camera phone based currency reader for the visually impaired. In Proceedings of the 10th international ACM SIGACCESS Conference on Computers and accessibility (Assets '08). Association for Computing Machinery, New York, NY, USA, pp. 305-306. https://doi.org/10.1145/1414471.1414551

[8] Neumann, L., Matas, J. (2013). Scene text localization and recognition with oriented stroke detection. IEEE International Conference on ComputerVision, pp. 97-104. https://doi.org/10.1109/ICCV.2013.19

[9] Field, L. (2014). Improving text recognition in images of natural scenes. Ph.D. dissertation. Computer Science, Univ. Massachusetts Amherst, Amherst, MA, USA.

[10] Weinman, J.J.,Learned-Miller, E., Hanson, A.R. (2009). Scene text recognition using similarity and a lexicon with sparse belief propagation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 31(10): 1733-1746. https://doi.org/10.1109/TPAMI.2009.38

[11] Jung, K., Kim, K.I., Jain, A.K. (2004). Text information extraction in images and video: A survey. Pattern Recognition, 37(5): 977-997. https://doi.org/10.1016/j.patcog.2003.10.012

[12] Yin, X., Zuo, Z., Tian,S., Liu, C. (2016). Text detection, tracking and recognition in video: A comprehensive survey. IEEE Transactions on Image Processing, 25(6): 2752-2773. https://doi.org/10.1109/TIP.2016.2554321

[13] Zhong, Y., Zhang, H.J., Jain, A.K. (2000). Automatic caption localization in compressed video. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(4): 385-392. https://doi.org/10.1109/34.845381

[14] Yin, X., Yin, X., Huang, K., Hao, H. (2014). Robust text detection in natural scene images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 36(5): 970-983. https://doi.org/10.1109/TPAMI.2013.182

[15] Dalton, J., Allan, J., Mirajkar, P. (2013). Zero-shot video retrieval using content and concepts. In Proceedings of the 22nd ACM International Conference on Information & Knowledge Management (CIKM '13). Association for Computing Machinery, pp. 1857-1860. https://doi.org/10.1145/2505515.2507880

[16] Lucas, S.M., Panaretos, A., Sosa, L., Tang, A., Wong, S., Young, R. (2003). ICDAR 2003 robust reading competitions. Seventh International Conference on Document Analysis and Recognition, Proceedings, pp. 682-687. https://doi.org/10.1109/ICDAR.2003.1227749

[17] Wang, K., Babenko, B., Belongie, S. (2011). End-to-end scene text recognition, 2011 International Conference on Computer Vision, pp. 1457-1464. https://doi.org/10.1109/ICCV.2011.6126402

[18] Wang, K., Belongie, S. (2010). Word Spotting in the Wild. Computer Vision – ECCV 2010, pp. 591-604. https://doi.org/10.1007/978-3-642-15549-9_43

[19] Jung, J., Lee, S., Cho, M.S., Kim, J.H. (2011). Touch TT: Scene text extractor using touchscreen interface. ETRI Journal, 33: 78-88. https://doi.org/10.4218/ETRIJ.11.1510.0029

[20] Yao, C., Bai, X., Liu, W., Ma, Y., Tu, Z. (2012). Detecting texts of arbitrary orientations in natural images. 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1083-1090. https://doi.org/10.1109/CVPR.2012.6247787

[21] Mishra, A., Alahari, K., Jawahar, C. (2012). Scene text recognition using higher order language priors. In Proceedings British Machine Vision Conference 2012, pp. 127.1-127.11. https://doi.org/10.5244/C.26.127

[22] Karatzas, D., Shafait, F., Uchida, S. (2013). ICDAR 2013 robust reading competition. 2013 12th International Conference on Document Analysis and Recognition, pp. 1484-1493. https://doi.org/10.1109/ICDAR.2013.221

[23] Karatzas, D., Gomez-Bigorda, L., Nicolaou, A., et al. (2015). ICDAR 2015 competition on robust reading. 2015 13th International Conference on Document Analysis and Recognition (ICDAR), pp. 1156-1160. https://doi.org/10.1109/ICDAR.2015.7333942

[24] Veit, A., Matera, T., Neumann, L., Matas, J., Belongie, S. (2016). Coco-text: Dataset and benchmark for text detection and recognition in natural images. arXiv preprint arXiv:1601.07140.

[25] Gupta, A., Vedaldi, A., Zisserman, A. (2016). Synthetic data for text localisation in natural images. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2315-2324. https://doi.org/10.1109/CVPR.2016.254

[26] Liu, Y.L., Jin, L.W., Zhang, S.T., Luo, C.J., Zhang, S. (2019). Curved scene text detection via transverse and longitudinal sequence connection. Pattern Recognition, 90: 337-345. https://doi.org/10.1016/j.patcog.2019.02.002

[27] Nayef, N., Yin, F., Bizid, I. et al. (2017). ICDAR2017 robust reading challenge on multi-lingual scene text detection and script identification - RRC-MLT. 201714th

IAPR International Conference on Document Analysis and Recognition (ICDAR), pp. 1454-1459. https://doi.org/10.1109/ICDAR.2017.237

[28] Shi, B., Yao, C., Liao, M., Yang, M., Xu, P., Cui, L., Bai, X. (2017). ICDAR2017 competition on reading Chinese text in the wild (RCTW-17). 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), pp. 1429-1434. https://doi.org/10.1109/ICDAR.2017.233

[29] Ch'ng, C.K., Chan, C.S. (2017). Total-text: A comprehensive dataset for scene text detection and recognition. 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), pp. 935-942, https://doi.org/10.1109/ICDAR.2017.157

[30] Sun, Y.P., Liu, J.M., Liu, W., Han, J.Y., Ding, E., Liu, J.T. (2019). Chinese street view text: Large-scale Chinese text reading with partially supervised learning. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), pp. 9085-9094. https://doi.org/10.1109/ICCV.2019.00918

[31] Nayef, N., Patel, Y., Busta, M. et al. (2019). ICDAR2019 robust reading challenge on multi-lingual scene text detection and recognition — RRC-MLT-2019. pp. 1582-1587. https://doi.org/10.1109/ICDAR.2019.00254

[32] Chng, C.K., Liu, Y.L., Sun, Y.P., et al. (2019). ICDAR2019 robust reading challenge on arbitrary-shaped text - rrc-art. 2019 International Conference on Document Analysis and Recognition (ICDAR), pp. 1571-1576. https://doi.org/10.1109/ICDAR.2019.00252

[33] Smeaton, A.F., Over, P. (2002). The TREC-2002 video track report. In Proc. 17th Text Retr. Conf. (TREC), pp. 1-17.

[34] Xu, J., Mei, T., Yao, T., Rui, Y. (2016). MSR-VTT: A large video description dataset for bridging video and language. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5288-5296. https://doi.org/10.1109/CVPR.2016.571

[35] Merino-Gracia, C., Mirmehdi, M. (2014). Real-time text tracking in natural scenes. IET Comput. Vis., 8(6): 670-681. https://doi.org/10.1049/iet-cvi.2013.0217

[36] Nguyen, P.X., Wang, K., Belongie, S. (2014). Video text detection and recognition: Dataset and benchmark. Proc. IEEE Winter Conf. Appl. Comput. Vis., pp. 776-783. https://doi.org/ 10.1109/WACV.2014.6836024

[37] Reddy, S., Mathew, M., Gomez, L., et al. (2020). RoadText-1K: Text detection & recognition dataset for driving videos. IEEE International Conference on Robotics and Automation (ICRA), pp. 11074-11080. https://doi.org/10.1109/ICRA40945.2020.9196577

[38] Jung, K., Kim, K.I., Jain, A.K. (2004). Text information extraction in images and video: A survey. Pattern Recognition, 37(5): 977-997. https://doi.org/10.1016/j.patcog.2003.10.012

[39] Lee, S.H., Cho, M.S., Jung, K., Kim, J.H. (2010). Scene text extraction with edge constraint and text collinearity. 2010 20th International Conference on Pattern Recognition, pp. 3983-3986. https://doi.org/10.1109/ICPR.2010.969

[40] Kim, K.I., Jung, K., Kim, J.H. (2003). Texture-based approach for text detection in images using support vector machines and continuously adaptive mean shift algorithm. IEEE Trans. Pattern Anal. Mach. Intell., 25(12): 1631-1639. https://doi.org/10.1109/TPAMI.2003.1251157

[41] Guan, L.B., Chu, J.Z. (2017). Natural scene text detection based on SWT, MSER and candidate classification. 2nd International Conference on Image, Vision and Computing (ICIVC), pp. 26-30. https://doi.org/10.1109/ICIVC.2017.7984452

[42] Gupta, N., Jalal, A.S. (2019). A robust model for salient text detection in natural scene images using MSER feature detector and Grabcut, Multimedia Tools and Applications, 78: 10821-10835. https://doi.org/10.1007/s11042-018-6613-1

[43] Wang, K., Babenko, B., Belongie, S. (2011). End-to-end scene text recognition. 2011 International Conference on Computer Vision, pp. 1457-1464. https://doi.org/10.1109/ICCV.2011.6126402

[44] Shelhamer, E., Long, J., Darrell, T. (2017). Fully convolutional networks for semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(4): 640-651. https://doi.org/10.1109/TPAMI.2016.2572683

[45] Liao, M., Shi, B., Bai, X., Wang, X., Liu, W. (2017). Textboxes: A fast text detector with a single deep neural network. Proc. AAAI Conf. on Artif. Intell.

[46] Gupta, A., Vedaldi, A., Zisserman, A. (2016). Synthetic data for text localisation in natural images. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2315-2324. https://doi.org/10.1109/CVPR.2016.254

[47] Redmon, J., Divvala, S., Girshick, R., Farhadi, A. (2016). You only look once: Unified, real-time object detection. Proc. IEEE Conf. on Comp. Vision and Pattern Recognit., pp. 779-788. https://doi.org/ 10.1109/CVPR.2016.91

[48] Ma, J., Shao, W., Ye, H., Wang, L., Wang, H., Zheng, Y., Xue, X. (2018). Arbitrary-oriented scene text detection via rotation proposals. IEEE Trans. on Multimedia, 20(11): 3111-3122. https://doi.org/10.1109/TMM.2018.2818020

[49] Ren, S., He, K., Girshick, K., Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. Proc. Adv. in Neural Info. Process. Syst., pp. 91-99. https://doi.org/10.1109/TPAMI.2016.2577031

[50] Liao, M., Shi, B., Bai, X. (2018). Textboxes++: A single-shot oriented scene text detector. IEEE Trans. on Image Process., 27(8): 3676-3690. https://doi.org/10.1109/TIP.2018.2825107

[51] Zhang, Z., Zhang, C., Shen, W., Yao, C., Liu, W., Bai, X. (2016). Multi-oriented text detection with fully convolutional networks. Proc. IEEE Conf. on Comp. Vision and Pattern Recognition, pp. 4159-4167. https://doi.org/ 10.1109/CVPR.2016.451

[52] Yao, C., Bai, X., Sang, N., Zhou, X., Zhou, S., Cao, Z. (2016). Scene text detection via holistic, multi-channel prediction. arXiv preprint arXiv:1606.09002.

[53] Wu, Y., Natarajan, P. (2017). Self-organized text detection with minimal post-processing via border learning. 2017 IEEE International Conference on Computer Vision (ICCV), pp. 5010-5019. https://doi.org/ 10.1109/ICCV.2017.535

[54] Liao, M.H., Lyu, P.Y., He, M.H., Yao, C., Wu, W.H., Bai, X. (2019). Mask TextSpotter: An end-to-end trainable neural network for spotting text with arbitrary shapes. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(2): 532-548. https://doi.org/10.1109/TPAMI.2019.2937086

[55] He, K., Gkioxari, G., Dollar, P., Girshick, R. (2017). Mask R-CNN. Proc. IEEE Int. Conf. on Comp. Vision, pp. 2980-2988. https://doi.org/10.1109/ICCV.2017.322

[56] Liao, M., Zhu, Z., Shi, B., Xia, G., Bai, X. (2018). Rotation sensitive regression for oriented scene text detection. Proc. IEEE Conf. on Comp. Vision and Pattern Recognition, pp. 5909-5918. https://doi.org/10.1109/CVPR.2018.00619

[57] Lyu, P., Yao, C., Wu, W., Yan, S., Bai, X. (2018). Multi-oriented scene text detection via corner localization and region segmentation. Proc. IEEE Conf. on Comp. Vision and Pattern Recognit., pp. 7553-7563. https://doi.org/10.1109/CVPR.2018.00788

[58] Liu, J., Liu, X., Sheng, J., Liang, D., Li, X., Liu, Q. (2019). Pyramid mask text detector. arXiv:1903.11800.

[59] Liao, M., Wan, Z., Yao, C., Chen, K., Bai, X. (2020). Real-time scene text detection with differentiable binarization. Proceedings of the AAAI Conference on Artificial Intelligence, 34: 11474-11481. https://doi.org/10.1609/aaai.v34i07.6812

[60] Liu, Y., Chen, H., Shen, C., He, T., Jin, L., Wang, L. (2020). ABCNet: Real-time scene text spotting with adaptive Bezier-curve network. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9806-9815. https://doi.org/10.1109/CVPR42600.2020.00983

[61] Wang, T., Wu, D.J., Coates, A., Ng, A.Y. (2012). End-to-end text recognition with convolutional neural networks. Proc. Int. Conf. on Pattern Recognit. (ICPR), pp. 3304-3308.

[62] Neubeck, A., Van Gool, L. (2006). Efficient non-maximum suppression. Proc. Int. Conf. on Pattern Recognit. (ICPR), 3: 850-855. https://doi.org/10.1109/ICPR.2006.479

[63] Bissacco, A., Cummins, M., Netzer, Y., Neven, H. (2013). PhotoOCR: Reading text in uncontrolled conditions. Proc. IEEE Int. Conf. on Comp. Vision, pp. 785-792. https://doi.org/10.1109/ICCV.2013.102

[64] Jaderberg, M., Simonyan, K., Vedaldi, A., Zisserman, A. (2016). Reading text in the wild with convolutional neural networks. Int. J. of Comp. Vision, 116(1): 1-20. https://doi.org/10.1007/s11263-015-0823-z

[65] Borisyuk, F., Gordo, A., Sivakuma, Ar. (2018). Rosetta: Large scale system for text detection and recognition in images. Proc. ACM SIGKDD Int. Conf. on Knowledge Discovery & Data Mining, pp. 71-79. https://doi.org/10.1145/3219819.3219861

[66] Wang, J., Hu, X. (2017). Gated recurrent convolution neural network for OCR. Proc. Adv. in Neural Inf. Process. Syst., pp. 334-343.

[67] Liu, W., Chen, C., Wong, K.K., Su, Z., Han, J. (2016). STAR-Net: A spatial attention residue network for scene text recognition. Proc. Brit. Mach. Vision Conf. (BMVC). BMVA Press, pp. 43.1-43.13. https://doi.org/10.5244/C.30.43

[68] Jaderberg, M., Simonyan, K., Zisserman, A., Kavukcuoglu, K. (2015). Spatial transformer networks. Proceedings of the 28th International Conference on Neural Information Processing Systems, 2: 2017-2025.

[69] Liu, W., Chen, C., Wong, K.K. (2018). Char-net: A character aware neural network for distorted scene text recognition. Proc. AAAI Conf. on Artif. Intell.

[70] Wang, Q., Huang, Y., Jia, W., He, X., Lu, Y., Blumenstein, M., Huang, Y. (2020). FACLSTM: ConvLSTM with focused attention for scene text recognition. Science China Information Sciences, 63: 120103. https://doi.org/10.1007/s11432-019-2713-1

[71] Dai, P., Zhang, H., Cao, X. (2021). SLOAN: Scale-adaptive orientation attention network for scene text recognition. IEEE Transactions on Image Processing, 30: 1687-1701. https://doi.org/10.1109/TIP.2020.3045602

[72] Nguyen, P.X., Wang, K., Belongie, S. (2014). Video text detection and recognition: Dataset and benchmark. IEEE Winter Conference on Applications of Computer Vision, pp. 776-783. https://doi.org/10.1109/WACV.2014.6836024

[73] Yilmaz, A., Javed, O., Shah, M. (2006). Object tracking: A survey. ACM Comput. Surv., 38(4): 13–es. https://doi.org/10.1145/1177352.1177355

[74] Lienhart, R., Effelsberg, W. (2000). Automatic text segmentation and text recognition for video indexing. Multimedia Syst., 8(1): 69-81. https://doi.org/10.1007/s005300050006

[75] Wang, B., Liu, C., Ding, X. (2013). A research on video text tracking and recognition. Proc. SPIE, vol. 8664, https://doi.org/10.1117/12.2009441

[76] Wang, L., Wang, Y., Shan, S., Su, F. (2018). Scene text detection and tracking in video with background cues. Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval, pp.160-168. https://doi.org/10.1145/3206025.3206051

[77] Horn, B.K.P., Schunck, B.G. (1981). Determining optical flow. Artificial Intelligence, 17(1-3): 185-203. https://doi.org/10.1016/0004-3702(81)90024-2

[78] Ji, S., Xu, W., Yang, M., Yu, K. (2013). 3D convolutional neural networks for human action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(1): 221-231. https://doi.org/10.1109/TPAMI.2012.59

[79] Shi, X., Chen, Z., Wang, H., Yeung, D.Y., Wong, W.K., Woo, W.C. (2015). Convolutional LSTM network: A machine learning approach for precipitation nowcasting. arXiv:1506.04214.

[80] Yu, H.Y., Huang, Y., Pi, L.H., Zhang, C.Q., Li, X., Wang, L. (2021). End-to-end video text detection with online tracking. Pattern Recognition, 113: 107791. https://doi.org/10.1016/j.patcog.2020.107791

[81] Mita, T., Hori, O. (2001). Improvement of video text recognition by character selection. Proceedings of Sixth International Conference on Document Analysis and Recognition, pp. 1089-1093. https://doi.org/10.1109/ICDAR.2001.953954

[82] Greenhalgh, J., Mirmehdi, M. (2015). Recognizing text-based traffic signs. IEEE Transactions on Intelligent Transportation Systems, 16(3): 1360-1369. https://doi.org/10.1109/TITS.2014.2363167

[83] Rong, X., Yi, C., Yang, X., Tian, Y. (2014). Scene text recognition in multiple frames based on text tracking. 2014 IEEE International Conference on Multimedia and Expo (ICME), pp. 1-6. https://doi.org/10.1109/ICME.2014.6890248

[84] Jain, M., Mathew, M., Jawahar, C. (2017). Unconstrained scene text and video text recognition for Arabic script. First International Workshop on Arabic Script Analysis and Recognition (ASAR). pp. 26-30. https://doi.org/10.1109/ASAR.2017.8067754

[85] Lu, W., Sun, H.B., Chu, J.H., Huang, X.D., Yu, J.X. (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: 40198-40211. https://doi.org/10.1109/ACCESS.2018.2851942

[86] Shi, B.G., Bai, X., Yao, C. (2015). An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(11): 2298-2304. https://doi.org/10.1109/TPAMI.2016.2646371

[87] Shi, B.G., Wang, X.G., Lyu, P.Y., Yao, C., Bai, X. (2016). Robust scene text recognition with automatic rectification. Proc. IEEE Conf. on Comp. Vision and Pattern Recognit., pp. 4168-4176. https://doi.org/10.1109/CVPR.2016.452

[88] Kasturi, R., Goldgof, D., Soundararajan, P., et al. (2009). Framework for performance evaluation of face, text, and vehicle detection and tracking in video: Data, metrics, and protocol. IEEE Transactions on Pattern Analysis and Machine Intelligence, 31(2): 319-336. https://doi.org/10.1109/TPAMI.2008.57

[89] Zhong, Y., Zhang, H.J., Jain, A.K. (2000). Automatic caption localization in compressed video. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(4): 385-392. https://doi.org/10.1109/34.845381

[90] Ye, Q., Huang, Q., Gao, W., Zhao, D. (2005). Fast and robust text detection in images and video frames. Image and Vision Comput., 23(6): 565-576. https://doi.org/10.1016/j.imavis.2005.01.004

[91] Haritaoglu, I. (2001). Scene text extraction and translation for handheld devices. Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001. https://doi.org/10.1109/CVPR.2001.990990

[92] Liang, J., Doermann, D., Li, H. (2005). Camera-based analysis of text and documents: A survey. Int. J. Doc. Anal. Recognit., 7: 84-104. https://doi.org/10.1007/s10032-004-0138-z

[93] Chen, X.L., Yang, J., Zhang, J., Waibel, A. (2004). Automatic detection and recognition of signs from natural scenes. IEEE Transactions on Image Processing, 13(1): 87-99. https://doi.org/10.1109/TIP.2003.819223

[94] He, Z.W., Liu, J.L., Ma, H.Q., Li, P.H. (2003). A new automatic extraction method of container identity codes. IEEE Trans. Intell. Transp. Syst., 2: 1688-1691. https://doi.org/10.1109/ITSC.2003.1252771

[95] Chen, D., Odobez, J.M., Bourlard, H. (2004). Text detection and recognition in images and video frames. Pattern Recognit., 37(3): 595-608. https://doi.org/10.1016/j.patcog.2003.06.001

[96] Mahajan, S., Rani, R. (2021) Text detection and localization in scene images: A broad review. Artif Intell Rev., 54: 4317-4377. https://doi.org/10.1007/s10462-021-10000-8

[97] Ye, Q.X., Doermann, D. (2015). Text detection and recognition in imagery: A survey. IEEE Trans. Pattern Anal. Mach. Intell., 37(7): 1480-1500. https://doi.org/10.1109/TPAMI.2014.2366765