International Journal on Document Analysis and Recognition (IJDAR) (2019) 22:143–162
https://doi.org/10.1007/s10032-019-00320-5

ORIGINAL PAPER

Scene text detection and recognition with advances in deep learning: a survey

Xiyan Liu[1,2] · Gaofeng Meng[1] · Chunhong Pan[1]

Received: 27 September 2017 / Revised: 26 February 2019 / Accepted: 6 March 2019 / Published online: 27 March 2019

Abstract
Scene text detection and recognition has become a very active research topic in recent several years. It can find many applications in reality ranging from navigation for vision-impaired people to semantic natural scene understanding. In this survey, we are intended to give a thorough and in-depth reviews on the recent advances on this topic, mainly focusing on the methods that appeared in the past 5 years for text detection and recognition in images and videos, including the recent state-of-the-art techniques on the following three related topics: (1) scene text detection, (2) scene text recognition and (3) end-to-end text recognition system. Compared with the previous survey, this survey pays more attention to the application of deep learning techniques on scene text detection and recognition. We also give a brief introduction of other related works such as script identification, text/non-text classification and text-to-image retrieval. This survey also reviews and summarizes some benchmark datasets that are widely used in the literature. Based on these datasets, performances of state-of-the-art approaches are shown and discussed. Finally, we conclude this survey by pointing out several potential directions on scene text detection and recognition that need to be well explored in the future.

Keywords Natural image · Text detection · Text recognition · Survey

1 Introduction

Text in natural scene image or video usually carries significant semantic information. Natural scene text detection and recognition aims to detect and locate text in scene image or video and recognize it automatically. It can find many applications, such as traffic monitoring, multimedia retrieval, semantic natural scene understanding.

Many techniques have been developed to detect and recognize the text in scene image and video. These techniques can be divided into three broad categories: text detection and localization, text recognition and end-to-end text recognition system. The goal of text detection and localization is to determine whether or not there is text in the given image or video and localize it [1–31]. Text recognition aims at converting the localized text in images into character coding [32–47], whereas the end-to-end text recognition system combines detection and recognition into a complete framework [48–55].

Scene text detection and recognition is a useful but challenging task. Figure 1 illustrates some challenging examples of scene text. In contrast to text recognition in documents, robust and accurate scene text detection and recognition is quite challenging due to many factors, including scene complexity, text diversity and more stringent practical requirements, which are analyzed as follows.

- Scene complexity: Images or videos generally suffer from noise, distortion, non-uniform illumination, partial occlusion, as well as confusion of the text and background. Complex background brings some obstacles to text detection or recognition in real world.
- Text diversity: Scene text vary in color, size, orientation, font, language, and text partial deletion, etc.

✉ Gaofeng Meng
gfmeng@nlpr.ia.ac.cn

Xiyan Liu
xiyan.liu@nlpr.ia.ac.cn

Chunhong Pan
chpan@nlpr.ia.ac.cn

[1] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, People's Republic of China

[2] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, People's Republic of China

Fig. 1 Examples of challenges of scene text detection. **a** Different orientations; **b** different languages; **c** different colors; **d** different size; **e** complex background; **f** occlusion; **g** blur; **h** noise; **i** non-uniform illumination

– More stringent practical requirements: Higher demands are often proposed including real-time processing and system robustness enhancement in many applications.

Recently, many good surveys of scene text detection and recognition have been published. We recommend that researchers can refer to these articles. For example, Ye et al. [56] elaborated and analyzed text detection and recognition from several aspects, including the facing challenges, fundamental problems and sub-problems, and some special issues related to text in video, multi-orientation and multilingual content. Zhu et al. [57] comprehensively surveyed a large number of algorithms for text detection and recognition. Yin et al. [58] mainly focused on text detection, tracking and recognition in video. In addition, many PhD theses also provided good surveys of this topic. Weinman [59] surveyed scene text reading from three parts: text detection, text recognition and joint detection and recognition. Field [60] summarized character-level and word-level detection and recognition. Jaderberg [61] has reviewed the related algorithms and datasets about the text spotting task. They also proposed two methods to generate training dataset. In addition, character-centric text spotting and word region proposal-based text detection were proposed. Mishra [62] detailed the challenges, problems and applications of scene text understanding. For scene text recognition, they proposed three methods, including a text segmentation (binarization) technique, an energy minimization framework and a holistic word recognition framework.

The above-mentioned surveys give a comprehensive review on the research achievements in the early years. However, many recent advances, typically the methods for the last 5 years, are not concluded and reviewed in the surveys. In this paper, we present a comprehensive survey of scene text detection and recognition in both image and video that have been published over the past 5 years. Compared with the previous survey, we pay more attention to deep learning-based methods. In addition, we introduce and analyze some other related works such as script identification, text/non-text classification and text-to-image retrieval. Descriptions of benchmark datasets are given as well as the overview of evaluation protocols. Based on these datasets, we discuss and compare the performance and efficiency of the state-of-the-art methods. Finally, we discuss some open issues and list potential future directions.

The organization of the remainder of this paper is as follows. We review some classical previous methods in Sect. 2. We detail the key approaches and techniques developed for scene text detection and recognition in recent 5 years in Sect. 3. In Sect. 4, some benchmark datasets and evaluation protocols are described briefly. The performance of some state-of-the-art methods is also compared. Section 5 gives the discussion and meanwhile points out some potential future work. Finally, Sect. 6 concludes the paper.

2 Summary of previous methods

From a methodological point of view, we review the previous methods from OCR style methods, machine learning style methods, energy minimization-based methods and deep learning-based methods.

Traditional optical character recognition (OCR) focuses on reading text from documents images, while photo-optical character recognition (OCR) is developed to read text from scene images, which is more challenging due to the uncontrolled conditions, complex backgrounds and variations in text size, color and font, etc. Benefiting from the progress in machine learning and large-scale language modeling, Bissacco et al. [63] proposed a photo-OCR system for text extraction from images. Three text detection approaches were adopted to generate candidate regions. They also trained a deep neural network for character classification. Experimental results demonstrate the superior performance of this method in text recognition. Neumann et al. [4] presented a real-time commercial OCR systems. In their systems, they employed an OCR stage that trained with synthetic fonts to recognize text. Furthermore, they advanced an OCR classifier

to label character regions in [53]. Lee et al. [45] proposed a lexicon-free photo-OCR system, which combined recursive CNN and RNN, and achieved state-of-the-art performance in many benchmark datasets.

In addition to OCR style approaches, energy minimization-based methods have achieved promising results in terms of scene text detection and recognition. Most of such methods are based on conditional random field (CRF) (such as [64,65]), Markov random field (MRF) [66], etc. Koo et al. [67] treated text lines extraction problem as an energy minimization problem which can handle the interference between text lines. Based on conditional random field (CRF), Mishra et al. [68] proposed an advanced energy minimization framework to recognize words. They minimized the energy function via unary and pairwise terms and finally got the optimal word for the image. Shi et al. [33] built a conditional random field (CRF) model for text recognition. One of the three text detection techniques that Bissacco et al. [63] used was based on energy minimization. They adopted graph cuts to labeling text and non-text by minimizing a Markov random field (MRF) with learned energy terms.

Many CRF-based or MRF-based methods can also be viewed as machine learning-based methods; for instance, Pan et al. [65] devised a supervised conditional random field (CRF) model combining unary component properties and binary contextual component to filter out non-text components.

Some methods detect and recognize scene text based on machine learning, such as support vector machine (SVM), random forest and clustering-based methods and so on. SVM is a widely used and effective classification algorithm. Neumann et al. [4] chose a SVM model with RBF kernel as classifier in their method. Lee et al. [36] used a linear SVM-based classifier to rank and select the most informative features for scene text recognition. Sharma et al. [69] adopted a Gaussian kernel SVM classifier to identify script and a hidden Markov model (HMM)-based method for word and character recognition. Based on HOG and CNN features, Turki et al. [20] also trained an SVM classifier to filter out the non-text elements. Kang et al. [27] proposed a higher-order correlation clustering (HOCC) for scene text detection. In addition, a structured SVM was used to learn the parameters of HOCC. Yin et al. [3] developed the single-link clustering algorithm, which can effectively group character candidates into text candidates. They also trained an AdaBoost classifier to determine whether the text candidates correspond to real text.

Even though the machine learning-based methods achieved competitive results in some challenging situations, it is not easy to extract more abstract features than deep learning. Reading scene text has greatly benefited from deep learning-based approaches, which will be introduced in the following section in detail.

3 Recent advances on scene text detection and recognition

Scene text detection and recognition consists of three major tasks, namely text detection and localization, text recognition and end-to-end text recognition system [50].

In this section, we will discuss and analyze the classical and advanced algorithms for text detection and recognition in both image and video according to the above three parts. Furthermore, the approaches based on deep learning have been discussed and compared in detail. In addition, we also discuss the relative works such as script identification, text/non-text classification, text binarization and text-to-image retrieval.

3.1 Text detection and localization

Text detection and localization aims at processing the input images or videos, detecting the presence of text and generating candidate text position. Existing methods for text detection and localization can be roughly categorized into three major groups: connected component (CC)-based methods, texture-based methods and deep learning-based methods.

3.1.1 Connected component (CC)-based methods

The connected component (CC)-based methods find and combine small components into a large component, then filter out non-text components by classifier, and finally extract text from image and combine it into text region. These methods have advantages of minor calculation and efficient. However, they are plagued by several limitations such as the inability to handle rotation, scale changes, complex backgrounds and other challenging cases. With regard to CC-based methods, the most representative methods are maximally stable extremal regions (MSERs) [1] and stroke width transform (SWT) [2].

Neumann et al. [1] used the maximally stable extremal regions (MSERs) algorithm to detect text. They first generate candidate region and then classify it with the already trained character and non-character classification. Finally, the model generates text line. The flowchart is shown in Fig. 2. Its novelty is embodied in adapt hypotheses verification framework to train multiple text lines and uses synthetic fonts to train the algorithm. MSER provides robustness to geometric and illumination conditions. However, this method only adapts to horizontal or nearly horizontal texts.

Epshtein et al. [2] proposed stroke width transform (SWT) method. SWT is a local image operator which uses the Canny edge detector [70] to perform edge detection and compute per pixel the width of the most likely stroke containing the pixel. SWT outputs an image, which contains the stroke width value in each pixel. Then, removing the interference of text candidate domain and clustering the characters that meet a

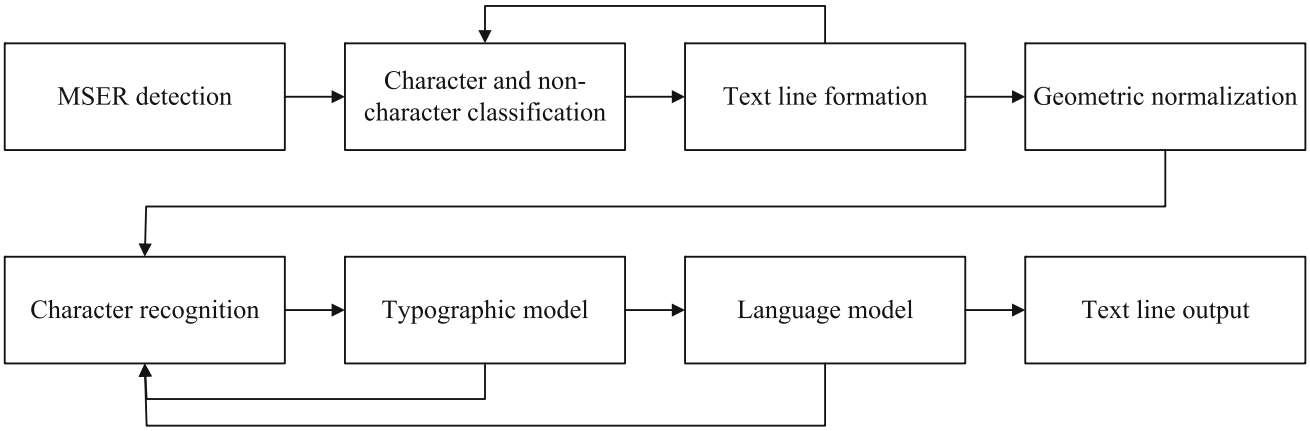

Fig. 2 Flowchart of the method presented by Neumann et al. [1]. This method introduced maximally stable extremal regions (MSERs) which provides robustness to geometric and illumination conditions

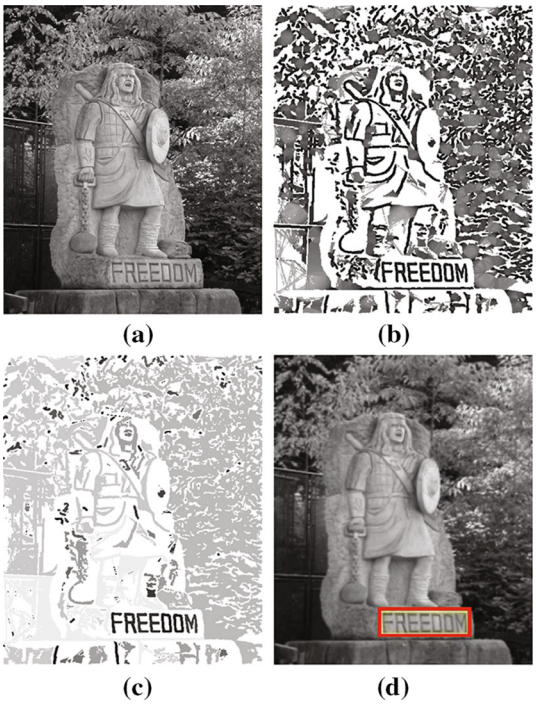

Fig. 3 Results of scene text detection by Epshtein et al. [2]. In their method, stroke width transform (SWT) is first introduced to distinguish text objects from non-textual objects from cluttered backgrounds

series of conditions, eventually, the text line is formed (see Fig. 3). Experiments show that SWT is highly efficient for text detection. This operator can detect texts in many fonts and languages, and it is insensitive to multi-scales and multi-directions. Nevertheless, SWT requires many human-defined constraints, so it may be failed in some challenging cases.

Yin et al. [3] developed MSER-based methods. They first extracted character candidates by the proposed MSERs pruning algorithm. Second, single-link clustering algorithm was adopted to cluster the character candidates into text candidates. Then, they trained a character classifier to eliminate non-text candidates. Finally, an AdaBoost classifier was used to detect text. However, there is room for further progress in detecting multi-orientation, multi-language or highly blurred texts in lower-resolution natural scene images.

The method proposed by Neumann et al. [4] treats the character detection problem as an efficient sequential selection from the set of extremal regions (ERs). This method takes up less memory, computes faster and maintains real-time performance. Similarly based on extremal regions (ERs), Cho et al. [5] presented an effective algorithm that can detect various texts. The algorithm extracted character candidates by extremal regions (ERs), and non-maximum suppression (NMS) was used to guarantee the uniqueness and compactness. In addition, double threshold and hysteresis tracking was adopted to fully detect texts even the candidates with low confidence. This method achieves high recall rate but is computationally expensive.

An efficient stroke detector was proposed by Busta et al. [6]. There are mainly three contributions. Firstly, stroke ending keypoint (SEK) and stroke bend keypoint (SBK) were introduced to detect stroke keypoint and then exploited to produce stroke segmentations. Secondly, they trained an AdaBoost-based classifier to classify text fragment and background clutter. Finally, based on text direction voting, they adopted a text clustering technique to group individual characters into text lines. It is worth noting that, besides computes fast, this method is scale- and rotation-invariant and supports a wide variety of scripts and fonts. However, it may be failed in some challenging cases, such as low image contrast, compact character.

3.1.2 Texture-based methods

The idea behind the texture-based method is that text in image has distinct textural properties, which can distinguish them

from the background. The techniques based on Gabor filters [71], Wavelet [72], fast Fourier transformation (FFT) [73], etc., can be used to detect the textural properties of a text region in an image [74].

In early years, Zhong et al. [7] proposed a filter that can directly detect text in the discrete cosine transform (DCT) domain. This algorithm runs fast but has a relatively low detection accuracy. Hanif et al. [8] proposed a cascade scene text detector and localizer. AdaBoost classifier was composed of linear discriminant classifier and likelihood ratio test (LRT). Following this work, in [9], they took into account the feature complexity in AdaBoost feature selection algorithm, which benefits the complexity of the strong classifier and the computational complexity of the real-time application. In addition, a neural network-based localizer was adopted to learn localization rules automatically. This method is suitable for text detection with various sizes and styles.

Zhang et al. [10] introduced a symmetry-based text line detector under the observation of the symmetry and self-similarity properties of character groups. This detector has the capacity to directly discover text lines from natural images. The model achieved state-of-the-art performance in ICDAR datasets and can be used for multi-language detection. However, its computational efficiency needs to be improved.

Liang et al. [11] proposed a new algorithm for arbitrary orientation character segmentation in videos and natural scene images based on wavelet decomposition. Laplacian Wavelet was also used to detect text candidate. Moreover, the horizontal and vertical sampling concept was proposed to segment characters from words. This method has advanced performance in terms of precision and F measure.

3.1.3 Deep learning-based methods

With the development of deep learning, convolutional neural network (CNN) has been widely explored. The main advantages of CNN are being insensitive to geometric transformation, deformation and illumination. It can extract information directly from image with a small computational cost. The performance of text detection and recognition in natural scene has been greatly improved due to the advanced properties of CNN. Existing CNN-based methods can be broadly categorized into several sub-classes: region proposal-based methods, segmentation-based methods and hybrid methods using multitask learning. Next, we will describe these methods in detail.

Region proposal-based methods take advantages of the region-based methods and the deep neural networks. These methods can robustly learn a component representation via CNN and can also make full use of CNN's powerful ability in classification and object detection. In the early years, Huang et al. [12] proposed a robust scene text detector combined maximally stable extremal regions (MSERs) with convolutional neural network (CNN). Experiments showed that this method achieved the best performance on ICDAR 2011. Combining faster R-CNN-based text detectors and LocNet-based localization module, Zhong et al. [13] proposed a technique to improve localization accuracy. This method can satisfy challenging cases and multi-language well.

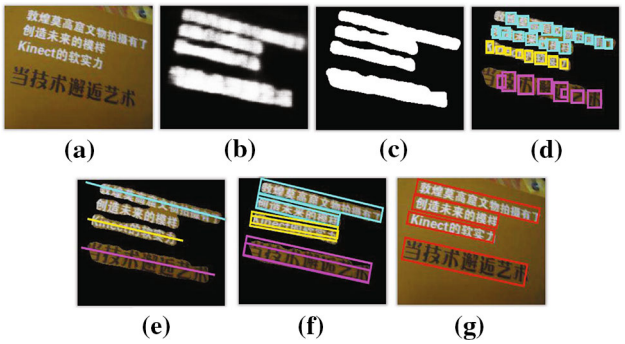

Fig. 4 The procedure of Zhang et al. [14]. **a** Input; **b** salient map of predicted text regions; **c** text block generation; **d** candidate character component extraction; **e** orientation estimation; **f** text line candidates extraction; **g** detection results

Zhang et al. [14] proposed a framework that can detect text in multiple orientations, languages and fonts (Fig. 4 shows the procedure). It mainly consists of two fully convolutional network (FCN): one of which is used for the salient map of text regions and the other FCN is used to predict the centroid of each character. However, false positives and missing characters may appear in certain situations, such as extremely low contrast, curvature, strong reflect light, too closed text lines or tremendous gap between characters. Another limitation is the low computational efficiency.

Zhu et al. [15] proposed a cascaded system for scene text detection. Text-Conv, a feature learning-based convolutional detector, was used to detect character patch. Region growing was applied to form CCs. The Word Graph algorithm groups detected character CCs into words based on their appearance and spatial relationships. It will achieve advanced performance by combining the above techniques together. However, the system failed to deal with some overlapped text lines, isolated characters, non-uniform illumination and so on.

Segmentation-based methods are generally adopted to produce more precise text regions. These methods are applied to multi-size text detection, but ineffective in detecting individual text lines or words. Segmentation network combined with detection network may obtain better performance. Guided by it, Qin et al. [16] focus on word-level text spotting and proposed a cascaded method with two convolutional neural networks. One is a fully convolutional network that called TextSegNet, which is designed to find text blocks. The other is a YOLO-like network named WordDetNet that detects

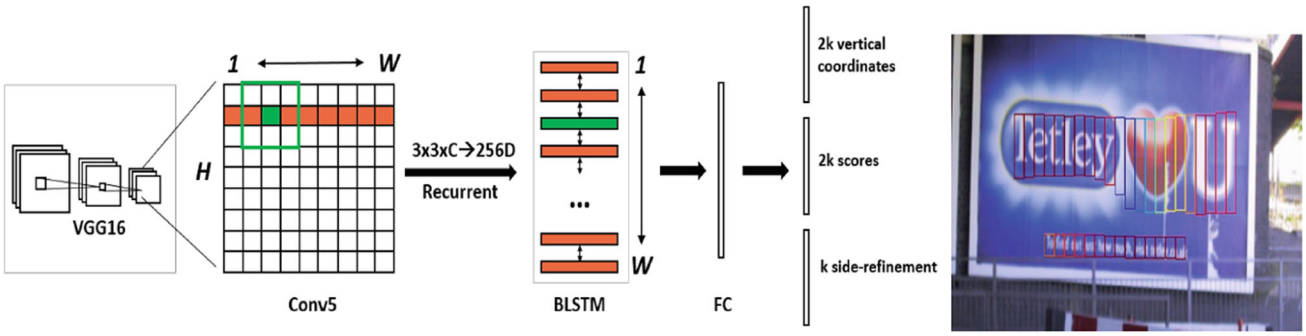

Fig. 5 Architecture of the CTPN that proposed by Tian et al. [22], which combines VGG16 and bidirectional LSTM

individual words by generating oriented rectangular regions. This method does a good job in improving F-score in ICDAR datasets.

Gupta et al. [17] developed a synthetic text scene image generation engine to generate images of natural scene text. Then, a large annotated dataset was built which called SynthText in the Wild. Inspired by fully convolutional networks (FCN), a new deep architecture was proposed, namely, fully convolutional regression network that was highly accurate, fast and trainable end-to-end.

The method proposed by Tang et al. [18] employed cascaded convolutional neural network(CNNs). In their work, three CNN-based models were proposed: detection network (DNet), segmentation network (SNet) and classification network (CNet). DNet was used to detect regions of text or the coarse candidate text region (CTR). SNet was effective in CTR refinement and text region segmentation. CNet was adopted to get the true text regions by classifying the refined CTRs into text or non-text. Benefiting from the power of deep CNN, this method achieves competitive precision and best recall and F measure in three benchmark datasets. However, the method is ineffective in the cases of non-uniform illumination, strong light and so on.

Based on superpixel segmentation and hierarchical clustering, Wang et al. [19] proposed a new character candidate extraction method. Inspired by the propriety of color consistency of the characters, they introduced a superpixel segmentation technique which integrated color and edge information together to output the refined superpixels. The character candidates were extracted via single-link clustering. In addition, they trained a deep convolutional neural networks (DCNN) classifier with double threshold strategy to classify text or non-text components. In the future, it can be extended to multiple connected characters detection task. *Hybrid methods* take into account both CC-based methods and texture-based methods. This class of methods usually employ multitask learning to optimize the model parameters. Turki et al. [20] presented an effective and robust scene text detector. They adopted Otsu method and a robust edge projection to filter the complex background. MSER method was utilized to detect a text pixels candidate. Based on CNN and HOG features, SVM was used to eliminate non-text elements. Finally, they eliminated false positive based on geometrical properties of text blocks.

Tian et al. [21] proposed a unified scene text detection system, termed as Text Flow. There are two key steps that constitute the system: One is character candidate detection that adopt cascade boosting technique. The other is text line extraction that uses a min-cost flow network. The min-cost flow network is able to integrate three sequential steps (false character candidate removal, text line extraction, text line verification) into a single process. Benefiting from the above advantages, this method eliminates the problem of error accumulation and achieves high recall. However, its speed has not been improved, and it cannot achieve a good detect results in some challenging conditions.

Tian et al. [22] presented a framework (as shown in Fig. 5) called Connectionist Text Proposal Network (CTPN) to localize text in the wild. CTPN is trained end-to-end and can be extended to multilingual and multi-scale text detection. This model achieves state-of-the-art results due to two key contributions. One is the vertical anchor mechanism which can improve localization accuracy by jointly predict location and text/non-text score of each text proposals. The other is the in-network recurrent neural network that is used to connect sequential text proposals; in this way, the model is able to deal with detection in most challenging cases.

He et al. [23] presented a novel text-attentional convolutional neural network (Text-CNN), which is effectively for extracting text-related regions and features from the image components. They proposed contrast-enhanced MSERs (CE-MSERs) detector for generating component candidates. The above two contributions effectively improve the detection accuracy and recall rate. However, it failed in some cases such as extremely ambiguous text.

Without plenty of assumptions like most existing methods, Fabrizio et al. [24] proposed a context-free text detection technique. Specifically, CC-based method was used to gen-

erate text candidates. Texture-based method was adopted to validate or discard the generated candidates. This method is applied to multi-sized, multi-oriented texts. However, false positives may occur due to fewer assumptions, and letters cannot be separated when they are stuck together.

3.2 Text recognition

Text recognition focuses on recognizing the text in text candidate under the assumption that the text has been successfully detected. Text recognition methods can be broadly divided into three categories: character-based methods, word-based methods and sequence-based methods. It is worth noting that the CNN+RNN framework for recognizing sequence-like text image after the first work [48] is quite popular with many advances to previous approaches.

3.2.1 Character-based methods

Character-based methods [32–34] perform character-level text recognition. As the basic element of text, character carries a large number of crucial information. Successful recognition of character makes bottom-up text recognition easier to implement.

Yao et al. [35] presented an advanced multi-scale representation for scene text recognition, called strokelets whose essence is a set of multi-scale mid-level primitives and can be automatically learned from bounding box labels. The main advantages of strokelets are multi-folds and can be summarized as usability, robustness, generality and expressivity. This method is effective at describing characters and robust in recognition. In [36], Lee et al. were dedicated to character recognition and presented a discriminative feature pooling method (see Fig. 6). Firstly, they extracted low-level features that include gradient histograms, gradient magnitude and color. Secondly, the extracted low-level features were integrated automatically via region-based feature pooling technique. Finally, a linear SVM-based classifier was devised to rank and select the most informative features. The proposed feature has proven to be compact, computationally efficient and can effectively model distinctive spatial structures of each individual character.

Lou et al. [37] devised a generative shape model for scene text recognition. This method achieves state-of-the-art results on ICDAR datasets and SVT datasets with very little training data. In order to handle the challenging case of variation of fonts in the wild, they adopted a greedy approach for representative fonts selection. Furthermore, they chose max-margin structured output learning to train a parsing model that is able to infer the true word. However, this method may miss edge evidence in the case of blur and overexposure.

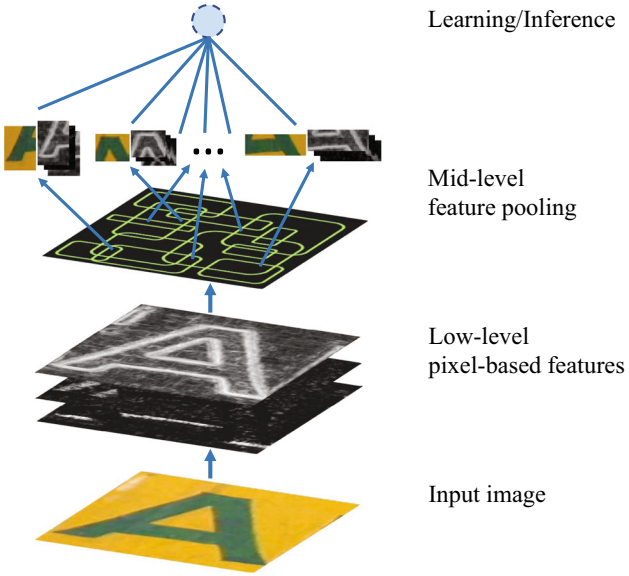

Fig. 6 Framework of region-based feature pooling algorithm proposed by Lee et al. [36]. The method incorporates pixel-wise low-level image features and subregion mining schema

3.2.2 Word-based methods

Word-based methods [38,39,49] recognize text at word level also attract the attention of many scholars. Some existing methods treat word recognition as an optimization problem, such as Phan et al. [40]; in order to recognize perspective scene texts of arbitrary orientations, they proposed an advanced text recognition method. By using dense SIFT in a bag-of-keypoints framework, character could be recognized robustly. With regard to word recognition, they adopted an optimized alignment algorithm. It is remarkable that the whole system was trained with only frontal character, which greatly reduced the demand for datasets.

Weinman et al. [41] proposed a reading system. They integrated word segmentation with recognition in the probabilistic framework. Lexical decision and sparse beam search tools were used to improve the recognition accuracy and operation efficiency.

3.2.3 Sequence-based methods

Recently, numerous works [42,43] have taken the problem of text recognition as a sequence recognition task, in which the text is represented via character sequence. Sequence-based methods consider the intimately relationship between characters. Shi et al. [44] proposed a method for irregular text recognition which is called RARE (Robust text recognizer with Automatic REctification). This model combined spatial transformer network (STN) and a sequence recognition network (SRN), where STN transformed an input image to a rectified image and SRN was applied to recognize text. In

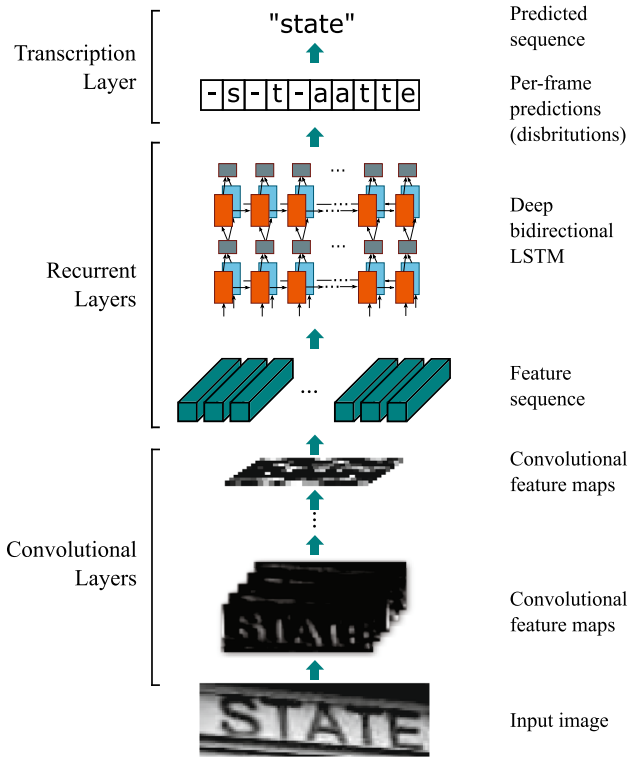

Fig. 7 The architecture of CRNN proposed by Shi et al. [48]. The model mainly consists of convolutional layers, recurrent layers and transcription layer

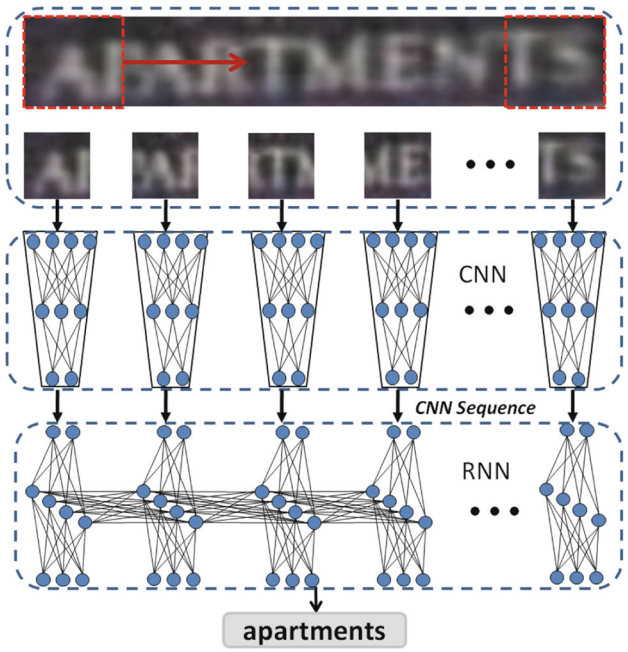

Fig. 8 The architecture of DTRN proposed by He et al. [46], which is a CNN + RNN structure

addition, attention-based model was used to extend the STN framework as well as SRN. This method can greatly improve the recognition rate of irregular text but tends to fail in some challenging cases, such as text with heavy perspective distortion or large curve angles.

Lee et al. [45] proposed a lexicon-free photo-OCR system called recursive recurrent neural networks with attention modeling (R2AM). In this paper, recursive CNN was used to capture image feature due to its ability of increasing the depth of the CNN and actively respond to the extracted features. RNN was adopted to model the character-level statistics of text. Furthermore, in order to force the model to pay more attention to the most important segments of incoming features, they chose soft-attention model in their method.

As for image-based sequence recognition problem, Shi et al. [48] proposed an end-to-end text recognition system, called convolutional recurrent neural network (CRNN)(see Fig. 7), which consists of convolutional layers, recurrent layers and transcription. By combining CNN with RNN, CRNN can handle sequences in arbitrary lengths and does not rely on additional lexicon. Furthermore, this method achieved competitive performance with less parameters.

He et al. [46] devised a deep-text recurrent network (DTRN) (see Fig. 8) that treats scene text recognition as a deep sequence labeling problem. This model adopted the structure of CNN + RNN, where CNN was applied to extract the features of objects and learn high-level image representation, RNN was modeled for sequence learning. In addition, this model does a good job in processing unknown words, arbitrary strings and multiple words without any pre-defined dictionary.

Yang et al. [47] focused on irregular text recognition task and presented an end-to-end model that consists of an auxiliary dense character detection model and an attention model. The former is a FCN architecture that is used to learn the high-level visual representations. The latter is guided by an alignment loss. The model has achieved competitive results due to the combination of the above two components.

3.3 End-to-end text recognition system

End-to-end text recognition system is a combination of two tasks: text detection and text recognition. A complete framework is used to implement image input, text detection and localization, text recognition, and final output result.

Jaderberg et al. [50] proposed an end-to-end system for text spotting. A novel convolutional neural network (CNN) architecture was proposed to generate a series of saliency maps. The CNN-based classifier can process the entire image instead of cropped proposal. Furthermore, in order to acquire additional training data, they proposed an automatic data mining technique to generate word and character-level annotated data from Flickr. The method mentioned in this paper achieved state-of-the-art performance in many

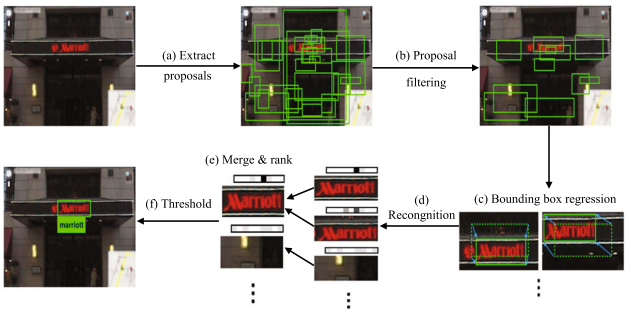

Fig. 9 Pipeline of the end-to-end text spotting method proposed by Jaderberg et al. [51]. **a** Region proposals extraction; **b** filtering false-positive detections; **c** refining proposals; **d** text recognition; **e** merging detections and give a score; **f** thresholding the detections results

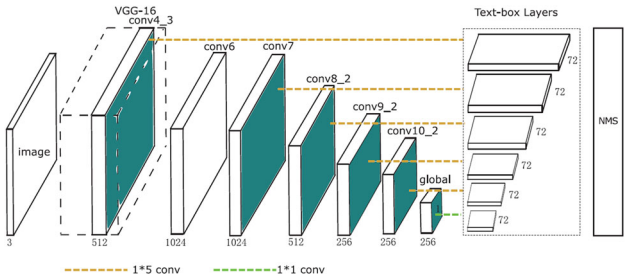

Fig. 10 TextBoxes architecture proposed by Liao et al. [55]. This model is a 28-layer fully convolutional network and achieved state-of-the-art performance of end-to-end text recognition

datasets. Lately, Jaderberg et al. [51] advanced text spotting methods through the following works. By combining edge boxes region proposal method with aggregate channel features (ACF) detector, the detection result showed competitive performance in recall rate. In addition, they devised a classifier to filter false-positive bounding boxes to reduce the computational costs and trained a regressor to refine the location of the bounding boxes. Inspired by the whole-word image algorithm, they trained a novel deep convolutional neural network with only synthetic data. The method is effective in many datasets, but it needs some artificial constraints. The pipeline is shown in Fig. 9.

Neumann et al. [49,52,53] have done a lot of researches and effectively promoted the progress in this field. In [49], with the advantages of sliding window and connected component methods, they proposed an unconstrained end-to-end real-time text localization and recognition method. The model detected strokes by oriented bar filters. Moreover, they also introduced a robust character representation. However, this method has many limitations in the detection process; for example, the subregion of a character might be another character, and the letter contains only one stroke on the boundary. In [52], they developed the end-to-end real-time system. In this literature, a region-based method was designed to detect initial text hypothesis in a single stage, followed by a more robust local text mode that was used to refine the text line hypothesis. In addition, they introduced a novel feature based on stroke support pixels (SSPs) which is invariant to scaling and rotations. Then, they made persistent efforts and proposed an advanced end-to-end real-time text localization and recognition method without any prior knowledge in [53]. With the advanced point of view that character detection problem can be treated as an efficient sequential selection from the set of extremal regions (ERs), real-time performance is achieved. This method is not affected by noise, low contrast, color variation and other interference factors.

Yao et al. [54] first proposed an approach on end-to-end recognition of multi-oriented texts. Benefitting from the discriminative power of the component level features and the inherent clustering mechanism of random trees, the component level classifier achieved competitive performance in text detection and recognition. Furthermore, a dictionary-based search method was designed to correct errors of recognition.

Recently, TextBoxes has been proposed by Liao et al. [55]. In this paper, they proposed an end-to-end model (as Fig. 10 shows) for text detection and recognition that achieved competitive results while keeping efficiency. A fully convolutional network was designed to detect scene text, and CRNN was adopted to recognize text. Experiments show that TextBoxes is the state-of-the-art of end-to-end text recognition and word spotting at present.

3.4 Text detection and recognition in scene video

Video text contains significant high level information which contributes to content-based video understanding and retrieval [75]. However, compared with image text, video text detection and recognition is more challenging due to its display form which can be categorized as layered caption text, embedded caption text and scene text.

Embedded captions have good directivity and summary on the semantic information of video frames. In [76], Tian et al. proposed a unified framework for embedded captions. It is a Bayesian-based method which includes multi-components such as text tracking, tracking-based text detection and tracking text recognition.

With regard to scene text in video, the background is usually more complex and the frames have much lower resolution. Thus, text detection and recognition of video frames has not been effectively addressed with the traditional techniques for scene image. To overcome such problems and improve the performance of scene text detection and recognition in video, many approaches adopt spatial and temporal information.

Shivakumara et al. [77] presented a method for multi-oriented text detection in video. They used Fourier–Laplacian filtering to identify candidate text regions. They classified CC as simple CC and complex CC based on the number

of intersection points. Skeletonization was used to segment CCs into text lines. In order to eliminate false positive, text string straightness and edge density were used at the end. This method achieved higher recall and F measure. Liang et al. [38] proposed a new idea of convolving Laplacian with wavelet sub-bands at multi-level in the frequency domain. They adopted MSER and SWT to group candidate text pixels as text regions, followed by the mutual nearest neighbor clustering which was devised to cluster candidate text regions belonging to the same text line. This is a very effective method for arbitrary orientations text detection and achieves good accuracies in both scene image text detection and video text detection.

Based on CNN, Yousfi et al. [78] explored three deep learning models for Arabic video text feature extraction: Deep belief networks, which can learn the weights in a layer-wise manner, were exploited to learn character reconstruction; deep auto-encoders based on multilayer perceptron were also used to learn character reconstruction; convolutional neural networks were used to learn an adequate representation for character classification. Finally, a BLSTM-CTC network was used to predict the text transcription.

Sharma et al. [69] proposed a new technique for multilingual video text recognition. In this literature, a spatial pyramid matching (SPM)-based pooling scheme with SIFT descriptors and SVM classifier was applied for script identification, followed by a approach based on hidden Markov model (HMM) to perform word and character recognition. Finally, they used the lexicon-based post-processing of the word recognition results to improve the accuracy.

3.5 Multi-orientation text detection

In recent years, there are increasing research interests that focused on multi-oriented text detection, which appears to be the recent trend of text reading in the wild. Compared with horizontal or near-horizontal text which has been successfully detected via various methods [12,15], multi-oriented text detection tasks are more challenging and have much room for improvement. In order to improve the performance of the multi-oriented text detector, numerous methods [3,11,14,21,25] have been presented. In the early work, Yin et al. [26] first proposed a unified distance metric learning framework for adaptive hierarchical clustering. Then, a coarse-to-fine grouping algorithm for text candidates construction was applied, and an efficient scene text detection infrastructure was valid for multi-orientation scene text detection. This method performs excellent in multi-orientation text detection, but it is weak in detecting seriously blurred text in low-resolution images.

Yao et al. [79] proposed an effective technique for arbitrary orientations texts detection. In their method, SWT and other features such as component level features and chain level features were adopted. Two-level classification scheme was used to reduce the impact of manual parameters. Experimental results illustrate the superior performance of this method in multi-orientation text detection.

Based on MSER, Kang et al. [27] presented higher-order correlation clustering (HOCC) for text of arbitrary orientations. Specifically, inspired by spatial alignment and appearance consistency of MSER, weak hypotheses are first proposed. Then, they used HOCC to partition the MSERs into text line candidates. Finally, a texture classifier was used to filter out the non-text areas. In this paper, they creatively treated text detection as a graph partitioning problem and succeeded in detecting text with multiple orientations, languages and fonts. Nevertheless, the method requires some prior knowledge and assumptions.

Gomez et al. [28] presented a text-specific object proposals algorithm. Compared with existing object proposals methods, this method does a good job of improving recall rate while it only extracts a few thousand proposals. They integrated the proposed text proposals algorithm with other whole-word recognition models to achieve state-of-the-art performance in end-to-end scene word spotting task.

Zhou et al. [29] designed a fast, simple yet powerful framework for detecting multi-oriented and multi-quadrilateral shapes text. The detection pipeline consists of two stages: One is a fully convolutional network (FCN) model which is used to produce word or line level predictions, and the other is a process of non-maximum suppression (NMS) to yield detection results. This model outperforms previous state-of-the-art methods in performance while running much faster. However, the model is limited in detecting text with longer text regions and may miss word or incorrectly detect text due to limited training set size.

Shi et al. [30] proposed an advanced multi-oriented text detection method named segment linking (SegLink) (as Fig. 11 shows). This method is mainly composed of two

Fig. 11 Shi et al. [30] presented SgeLink. **a** Segments detection; **b** link connects two adjacent segments; **c** detection results; **d–f** SegLink can detect long lines of latin/non-latin text

elements: segments and links. The former is an oriented box that covers a part of a word or text line. The latter is used to connect a pair of adjacent segments and indicate that they belong to a whole word or text line. In addition, they designed a fully convolutional neural network to detect segments and links. Despite that there are a few limitations to the proposed model (e.g., two thresholds need to be set manually without adjusted automatically; character spacing will affect the detection performance), sufficient experiments have showed the competitive accuracy and efficiency of the method.

Based on CNN, Liu et al. [31] proposed deep matching prior network (DMPNet). In their method, quadrilateral sliding windows were used to roughly recall the text. Then, they proposed shared Monte Carlo method to compute polygonal overlapping areas. In particular, a novel sequential protocol and relative regression were adopted to allow quadrangle instead of rectangle to finely localize text. However, the sliding window was designed manually which may be affect the performance of the model.

3.6 Preprocessing and extension

There are some relative research tasks that play a crucial role in scene text reading system. In this action, we focus on script identification, text/non-text image classification, scene text binarization and vision and language.

3.6.1 Script identification

As an important step of scene text reading system, script identification aims at determining the language categories of the text. In some cases, because different scripts (such as Chinese and English) have strong distinctions in stroke structure, many existing text detection and recognition techniques cannot run effectively on multi-scripts. However, some scripts have strong inter-class similarity, which is not conducive to the distinction, resulting in more challenges for the scene text reading system. Therefore, some methods contribute to the development of script recognition.

Nicolaou et al. [80] proposed an advanced script identification method which is applied to video text, scene text and handwritten text. This method integrates hand-crafted texture features with artificial neural network. Experiments showed that the model achieved state-of-the-art in CVSI 2015 Video Script Identification datasets and SIW dataset, etc.

Shi et al. [81] developed a technique for script identification in the wild. They chose deep network as the basic framework and proposed discriminative convolutional neural network (DisCNN) which has the ability to distinguish scripts with subtle differences. In their method, deep features with discriminative mid-level representations were combined to train the deep network. In addition, they gathered a large script dataset called SIW-13, which contains 16291 images cover 13 scripts in the wild. This method showed competitive performance in many datasets while it sometimes failed in challenging cases (e.g., blurred or low-resolution images, unusual text layouts and text with similar appearance).

Gomez et al. [82] presented a CNN-based method which is able to learn discriminative stroke-part representations and their relative importance in a patch-based classification scheme via conjoined convolutional networks. More specifically, patch-based classification scheme was designed to deal with the problem of extremely variable aspect ratio of text while identifying script. The model was evaluated in three benchmark datasets and outperformed majority of previous methods.

In addition, we also introduce several widely used datasets for script identification which cover scene text and video text. The details of CVSI 2015 [83] and SIW-13 [81] are given in Table 1.

3.6.2 Text/non-text classification

The goal of Text/non-text classification is to pick out images containing text from a large number of images. It can be considered as an inevitable preprocessing of the text detection and recognition. This classification is challenging due to the randomness of the text appearance, the complex backgrounds and the special requirements for time complexity. Several existing methods have promoted the development of this field; for example, Delaye et al. [84] proposed a conditional random fields-based classification for handwritten text. Phan et al. [85] also presented a RNN-based classifier to handle text/non-text classification in handwritten documents. In addition, Sharma et al. [86] focused on text/non-text classification in video.

Bai et al. [87] focused on text/non-text scene image classification. They proposed a CNN-based technique called multi-scale spatial partition network (MSP-Net). In this method, they predicted whether the image has text content on block level. As long as there exists text in one block of the image, it was considered a text image; otherwise, it was a non-text image. An effective and robust neural network architecture was designed and achieved good performance while it may fail in some challenging cases, such as low illumination, regular curves and so on.

3.6.3 Text binarization

Binarization also plays an important role in the preprocessing of text detection and recognition. Document image binarization aims at separating text from background, which contributes to the document images analysis and recognition. The methods of document image binarization can be broadly divided into three categories: threshold methods [88,89], clustering-based methods [90] and energy mini-

Table 1 Details of benchmark datasets

Dataset	Content	Task
ICDAR 2003 [100]/2005 [101]	Contains 258 training samples and 251 testing samples	For horizontal text detection and recognition
ICDAR 2011 [102]	Contains 484 natural and born-digital images	Extension of ICDAR 2003
ICDAR 2013 [103]	Contains 229 training samples and 233 testing samples	For horizontal text detection, recognition, and end-to-end recognition
ICDAR 2015 [104]	Contains 1000 training samples and 500 testing samples	For detecting and recognizing text with arbitrary orientation
Street View Text [68] (SVT)	Contains 101 training images and 249 testing images which were captured from Google street view	For text detection and recognition
SVT-Perspective [40]	Contains 238 images which include 639 cropped words	For evaluating perspective text recognition
COCO-Text [105]	Contains 63686 images with 173589 annotations	A large challenging dataset for scene text detection and recognition
IIIT5K [106]	Contains 2000 training images and 3000 testing images with text in both natural scenes and born-digital images	For recognition of regular text
Chars74k [107]	Contains 7705 scene images	For text recognition evaluation
KAIST [108]	Contains 3000 images	For text detection and segmentation
MSRA-TD500 [79]	Contains 500 images which include indoor scenes and street views	The first attempt for oriented text detection
OSTD [109]	Consists of 88 images which include indoor scenes and street views	For multi-oriented text detection evaluation
Total-Text [110]	Contains 1555 scene images with 9330 annotated words, which include horizontal text, multi-oriented text and curved text	For curved text detection and recognition
SynthText in the Wild [17]	Contains 800,000 training images	A synthetic dataset used for training detectors
CTW-12k [111]	Contains 12,263 annotated images	For Chinese text localization and end-to-end recognition
Cute80 [112]	Contains 80 curved text images	For irregular text recognition
CVSI 2015 [83]	Contains 15 scripts. English, Hindi, Bengali, Oriya, Gujrathi, Punjabi, Kannada, Tamil, Telugu, Arabic, Malayalam, Chinese, Thai, Japanese, Korean	For video text script identification
SIW-13 [81]	Contains 13 scripts. 16291 images. Arabic, Cambodian, Chinese, English, Greek, Hebrew, Japanese, Kannada, Korean, Mongolian, Russian, Thai, Tibetan	For script identification in wild scenes

mization methods [91,92]. Document image binarization is very challenging due to the non-uniform shading artifacts, noises caused by paper aging, stains, bleed through and paper creases, etc. Recently, numerous methods have been made to address the challenges of document image binarization. Tensmeyer et al. [93] proposed an FCN-based technique for document image binarization. In addition, they found that relative darkness (RD) features treated as an additional input feature performed best.

Combining deep convolutional neural networks (DCNNs) with deep transposed convolutional neural networks (DTCNNs), Peng et al. [94] advanced document image binarization by a CNN-based model with the encoder–decoder structure. They also adopted a fully connected conditional random

field (CRF) to relabel the confidence maps to improve the quality of binarized document images. Meng et al. [95] proposed DCNN-based method for degraded document image binarization. In this paper, they devised an encoder–decoder network architecture to produce refined binarization results.

Based on Markov random field (MRF) framework, Wang et al. [66] proposed a binarization method which took into account the stroke features of text and produced seed pixels automatically. The highly confident cluster centers can be acquired via collecting diverse weight seeds.

3.6.4 Vision and language

Text in natural images usually increases the semantic information of the scene, so combining the text in the scene with the visual cues helps to image classification or retrieval. Various approaches are dedicated to the exploration of this field. In early years, Ha et al. [96] proposed a textual visual association model for text-to-image retrieval, which adopted hypernetwork (HN) model and crossmodal query expansion with the learned model.

Mishra et al. [97] presented a query-driven search method for text-to-image retrieval. In this paper, they first located text approximately in the image and indexed the database for a set of vocabulary words. Then, they computed scores based on the presence of characters. Finally, inverted index file was used to perform retrieval, while the results were re-ranked by spatial ordering and spatial positioning.

Karaoglu et al. [98] focused on fine-grained classification and business logo retrieval by combining the textual information and visual cues together. They proposed an unsupervised word-level proposal extractor, which contains two steps. The former was used to locate the text in the image, while the latter took word proposals as the input of the recognizer to generate the word-level representation. Moreover, they thought that recall rate is more conducive to fine-grained classification and logo retrieval.

Rong et al. [99] focused on text localization and retrieval in unambiguous scene. They proposed a recurrent dense text localization network (DTLN) to decode CNN features into a variable length set of text instance detections. Then, the outputs of the DTLN were fed into another proposed model named context reasoning text retrieval (CRTR), which was used to retrieve scene text instances via natural language. They also collected an unambiguous text instance retrieval dataset for evaluation.

4 Evaluations and comparisons

In this section, we mainly have the following three tasks. We list and analyze the widely used benchmark datasets, and then, the evaluation protocols are discussed. Finally, we further show and compare the performance of some advanced techniques in public datasets.

4.1 Benchmark datasets

The benchmark datasets play a crucial role in scene text detection and recognition. Table 1 provides a comprehensive list of existing benchmark datasets.

As one of the most popular competitions in recent years, "robust reading" competition provides ICDAR datasets (2003/05/11/13/15) [100–104]. ICDAR datasets are the commonly used datasets which include texts in both image and video. In addition, Street View Text dataset (SVT) [68], SVT-perspective [40], COCO-Text dataset [105], IIIT5K dataset [106], MSRA-TD500 [79], Chars74k [107], KAIST [108], OSTD [109], Cute80 [112], RCTW dataset [111] and Total-Text [26] have collected challenging text due to the complex background and the diversity of the text in orientation, color, scale and so on. The details of the mentioned datasets are given in Table 1.

4.2 Evaluation protocols

In order to promote the research of scene text detection and recognition tasks, some standardized evaluation protocols were proposed. In the following content, we will introduce some widely used evaluation protocols.

4.2.1 Evaluation protocols for Latin text

Evaluation protocol for text detection. From ICDAR robust reading competitions in 2003 [100], we can see that its evaluation system is based on the notions of precision and recall.

The best match $m(r, R)$ for a rectangle r in a set of Rectangles R is defined as:

$$m(r, R) = \max m_p\left(r, r'\right) | r' \in \mathbf{R} \tag{1}$$

Then, precision and recall are defined as:

$$\text{Precision} = \frac{\sum_{r_e \in \mathbf{E}} m(r_e, T)}{|E|} \tag{2}$$

$$\text{Recall} = \frac{\sum_{r_t \in \mathbf{T}} m(r_t, E)}{|T|} \tag{3}$$

where T is the set of ground truth and E represents the set of estimated rectangles.

The standard F measure f which is treated as weighted harmonic mean of the precision and recall figures, is adopted to evaluate the performance of text detection. It is worth noting that the parameter α is used to control the relative weights of precision and recall and is usually set as 0.5 to give them equal weight.

$$f = \frac{1}{\frac{\alpha}{\text{Precision}} + \frac{1-\alpha}{\text{Recall}}} \tag{4}$$

In the ICDAR robust reading competitions: ICDAR 2011 [102] and ICDAR 2013 [103], the method of Wolf et al. [113] is employed to evaluate scene text detection approaches. This evaluation metric takes into account one-to-one matching, one-to-many matching and many-to-one matching.

Precision, recall and F measure f are defined as:

$$\text{Precision} = \frac{\sum_i^N \sum_j^{|D^i|} M_D\left(D_j^i, G^i\right)}{\sum_i^N |D^i|} \tag{5}$$

$$\text{Recall} = \frac{\sum_i^N \sum_j^{|G^i|} M_G\left(G_j^i, D^i\right)}{\sum_i^N |G^i|} \tag{6}$$

$$f = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{7}$$

where $|D^i|$ and $|G^i|$ are the number of detection rectangles and ground-truth rectangles in the ith image, respectively. N is the total number of samples in the dataset. $M_D(D_j^i, G^i)$ and $M_G(G_j^i, D^i)$ are the matching scores for D^j and G^j. Their values are set to 1 for one-to-one matching, 0.8 for one-to-many matching and 0 for no matching [18].

Evaluation protocol for text recognition. In the competitions of ICDAR 2011 [102], ICDAR 2013 [103] and ICDAR 2015 [104], to evaluate the word recognition accuracy, they simply use the edit distance with equal cost of deletions, substitutions and insertions. They normalize the edit distance by the length of the ground-truth transcriptions and then calculate the normalized edit distance between the ground truth and the transcription.

Evaluation protocol for end-to-end system. Karatzas et al. [104] further improved the protocol. The evaluation strategy considers both the efficiency of text detection and the capacity of text recognition via precision, recall rate and F measure.

4.2.2 Evaluation protocols for non-Latin text

Evaluation protocol for MSRA-TD 500 [79]. Because the dataset contains horizontal and slant texts, the overlap ratio is defined as follows:

$$m(G, D) = \frac{A(G' \cap D')}{A(G' \cup D')} \tag{8}$$

where G' is the axis-aligned ground truth rectangle, D' is the axis-aligned estimated rectangle. $A(G' \cap D')$, and $A(G' \cup D')$ represents the intersection area and the union area of G' and D', respectively. When the angle between the predicted rectangle and the ground truth is less than $\pi/8$, and the overlap ratio is greater than 0.5, it can be considered that the detection is correct.

$$\text{Precision} = \frac{|\text{TP}|}{|E|} \tag{9}$$

$$\text{Recall} = \frac{|\text{TP}|}{|T|} \tag{10}$$

$$f = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{11}$$

where TP is the set of true positive detection, and E and T denote the estimated rectangles set and the ground truth set.

Evaluation protocol for RCTW [111]. As for text localization, AP is adopted as the primary evaluation metric. In addition, F measure is also calculated to be compatible with former competitions. It should be noted that the intersection-over-union (IoU) is calculated on polygons rather than rectangles. As for end-to-end recognition, average edit distance (AED) is calculated to evaluate method performance. Lower AED means better performance. In order to be compatible with other competitions, they also calculate a normalized measure as follows:

$$\text{NED}(s_1, s_2) = \frac{edit\_dist(s_1, s_2)}{\max(l_1, l_2)} \tag{12}$$

where NED is the normalized edit distance, s_1 and s_2 denote the text strings of a matching pair, and l_1 and l_2 are their text lengths. Then, the measure is calculated using the following:

$$1 - \frac{\sum_{i=1}^n \text{NED}(s_{i1}, s_{i2})}{n} \tag{13}$$

where n is the number of matching pairs.

4.3 Performance of algorithms

The performances of some state-of-the-art algorithms are given in Tables 2, 3, 4, 5, 6, 7, 8, 9 and 10. Each table gives the evaluation results of single datasets in scene text detection, recognition and end-to-end system.

The P, R, F appearing in the table indicate precision, recall and F measure, respectively. "50" and "1k" are lexicon sizes, "Full" denotes the combined lexicon of all images in the benchmarks, and "None" means lexicon-free. It is worth noting that "strong" means the lexicon containing 100 words specific to each image; "weak" means a lexicon with all words in the testing set; "generic" means a lexicon containing 90k words.

Performance of scene text detection. With the ICDAR dataset, the precision of current mainstream methods has exceeded 90%, while recall and F measure still have room for further improvement. We can see that the performance of most existing methods on COCO dataset and SVT is still not

Table 2 Performance of existing methods on ICDAR 2003 ([a]the method is not constrained by lexicon)

Algorithm	Year	50	Full	50k	None
Recognition					
Cheng[a] [114]	2017	99.20	97.30	–	94.20
Shi[a] [48]	2017	98.70	97.60	95.50	89.40
Shi[a] [44]	2016	98.30	96.20	94.80	90.10
Lee[a] [45]	2016	97.90	97.00	–	88.70
Jaderberg [51]	2016	98.70	98.60	93.30	93.10
Jaderberg[a] [115]	2015	97.80	97.00	93.40	89.60
Jaderberg[a] [50]	2014	96.20	91.50	–	–
Alsharif[a] [116]	2014	93.10	88.60	85.10	–
Su [42]	2014	92.00	82.00	–	–
Mishra [68]	2012	81.78	–	–	–

Algorithm	Year	50	Full	None
End-to-end				
Jaderberg [51]	2016	91.00	87.00	79.00
Jaderberg[a] [50]	2014	80.00	75.00	–
Wang[a] [117]	2011	68.00	61.00	–

Table 3 Performance of existing methods on ICDAR 2011 (the method is not constrained by lexicon)

Algorithm	Year	*P*	*R*	*F*
Detection				
Gupta [17]	2016	91.50	74.80	82.30
Tang [18]	2017	90.20	85.90	88.00
Li [118]	2017	89.20	81.70	85.10
Liao [55]	2017	89.00	82.00	86.00
Tian [22]	2016	89.00	79.00	84.00
Zhong [13]	2017	88.45	88.81	88.63
Huang [12]	2014	88.00	71.00	78.00
Tian [21]	2015	86.24	76.17	80.89
Neumann [49]	2013	79.30	66.40	72.30

Algorithm	Year	50	Full
Recognition			
Lee [36]	2014	88.00	77.00
Su [42]	2014	91.00	83.00
Shi [33]	2013	87.04	82.87

Algorithm	Year	None
End-to-end		
Jaderberg [51]	2016	77.00

satisfactory due to their extremely challenging. Numerous CNN-based methods are very competitive and achieved state-of-the-art performance in various benchmarks, e.g., Hu et al. [119], Zhu et al. [15] and Tang et al. [18]. In addition, con-

Table 4 Performance of existing methods on ICDAR 2013 ([a]the method is not constrained by lexicon)

Algorithm	Year	*P*	*R*	*F*
Detection				
Hu [119]	2017	93.34	87.53	90.34
Zhong [13]	2017	93.00	86.70	89.74
He [120]	2017	92.00	81.00	86.00
Gupta [17]	2016	92.00	75.50	83.00
He [121]	2017	89.00	86.00	88.00
Liao [55]	2017	89.00	83.00	86.00
Zhang [10]	2015	88.00	74.00	80.00
Shi [30]	2017	87.70	83.00	85.30

Algorithm	Year	None
Recognition		
Shi[a] [48]	2017	89.6
Cheng[a] [114]	2017	93.30
Lee[a] [45]	2016	90.00
Jaderberg [51]	2016	90.80
Jaderberg[a] [115]	2015	81.80

Algorithm	Year	Strong	Weak	Generic	None
End-to-end					
Liao[a] [55]	2017	91.00	89.00	84.00	
Busta[a] [122]	2017	89.00	86.00	77.00	–
Gomez[a] [28]	2017	81.16	79.49	–	68.54
Jaderberg [51]	2016	–	–	–	77.00

Table 5 Performance of existing methods on ICDAR 2015 ([a]the method is not constrained by lexicon)

Algorithm	Year	*P*	*R*	*F*
Detection				
Zhu [15]	2016	93.39	81.02	86.77
Tang [18]	2017	91.90	87.10	89.50
Li [118]	2017	91.40	80.50	85.60
Wu [123]	2017	91.00	78.00	84.00
Zhou [29]	2017	83.27	78.33	80.72
He [120]	2017	82.00	80.00	81.00

Algorithm	Year	None
Recognition		
Cheng[a] [114]	2017	85.30

Algorithm	Year	Strong	Weak	Generic	None
End-to-end					
Li [118]	2017	91.08	89.81	84.59	–
Busta[a] [122]	2017	54.00	51.00	47.00	–
Gomez[a] [28]	2017	53.30	49.61	–	47.18
Neumann[a] [53]	2016	35.00	19.90	15.60	–

Table 6 Performance of existing methods on IIIT5K ([a]the method is not constrained by lexicon)

Algorithm	Year	50	1k	None
Recognition				
Cheng[a] [114]	2017	99.30	97.50	87.40
Yang[a] [47]	2017	97.80	96.10	–
Shi[a] [48]	2017	97.60	94.40	78.20
Jaderberg [51]	2016	97.10	92.70	–
Lee[a] [45]	2016	96.80	94.40	78.40
Shi[a] [44]	2016	96.20	93.80	81.90
Jaderberg[a] [115]	2015	95.50	89.60	–
Gordo [124]	2015	93.27	86.57	–
Almazan [125]	2014	88.57	75.60	–

Table 7 Performance of existing methods on SVT ([a]the method is not constrained by lexicon)

Algorithm	Year	*P*	*R*	*F*
Detection				
Tian [22]	2016	68.00	65.00	66.00
Zhang [10]	2015	68.00	53.00	60.00
Rong [99]	2017	29.00	27.00	28.00
Gupta [17]	2016	26.20	27.40	26.70
Tang [18]	2017	–	76.20	–

Algorithm	Year	50	None
Recognition			
Cheng[a] [114]	2017	97.10	85.90
Shi[a] [48]	2017	96.40	80.80
Lee[a] [45]	2016	96.30	80.70
Shi[a] [44]	2016	95.50	81.90
Jaderberg [51]	2016	95.40	80.70
Yang[a] [47]	2017	95.20	–
Gordo [124]	2015	91.81	–
Jaderberg[a] [115]	2015	93.20	91.70
Jaderberg[a] [50]	2014	86.10	–
Almazan [125]	2014	87.01	–
Su [42]	2014	–	83.00
Mishra [68]	2012	–	73.26

Algorithm	Year	50	None
End-to-end			
Gomez[a] [28]	2017	85.00	54.00
Jaderberg [51]	2016	82.00	57.00
Jaderberg[a] [50]	2014	56.00	–

nected component also plays a key role in this filed, and Yin et al. [3], Neumann et al. [4] and other CC-based methods also work well.

Table 8 Performance of existing methods on MSRA-TD500

Algorithm	Year	*P*	*R*	*F*
Detection				
Zhou [29]	2017	87.28	67.43	76.08
Shi [30]	2017	86.00	70.00	77.00
Zhang [14]	2016	83.00	67.00	74.00
Wu [123]	2017	77.00	78.00	77.00
He [120]	2017	77.00	70.00	74.00
Turki [20]	2017	72.00	79.00	75.33
Kang [27]	2014	71.00	62.00	66.00

Table 9 Performance of existing methods on COCO

Algorithm	Year	*P*	*R*	*F*
Detection				
Zhou [29]	2017	50.39	32.40	39.45
He [121]	2017	46.00	31.00	37.00
Hu [119]	2017	45.20	30.90	36.80

Table 10 Performance of methods on multilingual

Algorithm	Year	*P*	*R*	*F*
Detection				
Cho [5]	2016	93.10	93.50	93.30
Tian [21]	2015	84.70	78.40	81.40
Tian [22]	2016	84.00	80.00	82.00
Zhong [13]	2017	82.45	84.23	83.33

In addition, we also introduce and analyze the efficiency of previous state-of-the-art approaches in Fig. 12. Zhou et al. [29], Liao et al. [55], Shi et al. [30], etc., detect texts very fast. While ensuring the performance of the model, they greatly reduce the calculation time, which is beneficial to the model for the real application.

Performance of scene text recognition. CNN-based methods proposed by Cheng et al. [114] and Jaderberg et al. [51] show the best performance in ICDAR 2003, SVT and IIIT5K. The approach which combines the TextBoxes and CRNN [55] achieved the state-of-the-art performance of text spotting and end-to-end text recognition. Moreover, CNN+RNN framework is quite popular with many previous approaches [46,48].

Performance of end-to-end system. Gomez et al. [28] performed better than other techniques in SVT, IC03-50 and IC03-Full. Liao et al. [55] proposed TextBoxes and achieved the best performance in ICDAR 2013. Similarly, CNN-based methods occupy an important position.

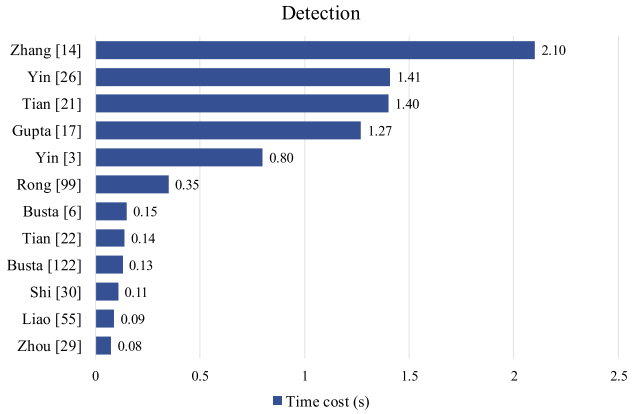

Fig. 12 Efficiency of existing methods in scene text detection

5 Discussion

We have made a fundamental survey on scene text detection and recognition from a methodological point of view. Numerous approaches have witnessed tremendous efforts and progresses in recent years. With respect to text detection and localization, the methods can be roughly classified into the following categories: connected component (CC)-based methods, texture-based methods and deep learning-based methods. Among them, the most representative methods of CC-based methods are MSER [1] and SWT [2], which are the basis of various state-of-the-art methods. With the rapid development of deep learning, a large amount of approaches [30,31,55,114,122] adopt neural networks and achieve competitive performance. In addition, multi-oriented text detection has attracted more interests, because it is more challenging and practical. As for scene text recognition, CNN+RNN framework [48] is quite popular with many advances to previous approaches. Moreover, many approaches [42,43,45,46] take the problem of text recognition as a sequence recognition task. Compared with text detection or recognition, end-to-end system is more challenging but has direct practical value. Last but not least, other contributions such as script identification, text binarization, text/non-text classification and text-to-image retrieval also play a crucial role in text reading system.

We discuss several potential trends in this field:

- Complex scene and large-scale dataset: As shown in Table 9, previous methods work poorly in COCO dataset due to its complex background and diverse text. Therefore, to build a system that can satisfy multiple challenging cases will become the research trend.
- Multilingual detection and recognition: In some cases, various languages may appear in the same scenario. Because different languages have strong distinctions in stroke structure (e.g., Chinese and English), plenty of existing techniques cannot work well on multilingual detection and recognition. Therefore, on the one hand, we can perform script identification firstly and then form specific model to detect and recognize each language separately. On the other hand, there are some techniques; for example, Text flow [21] can detect multilingual via a unified model.
- Real-time detection and recognition: With the explosive growth of mobile device, processing video data in real time are becoming an increasingly demand. In this task, the key issue is to optimize the running speed and save storage, and then establish an advanced end-to-end system to improve real-time performance, accuracy and robustness. Of course, several other applications need real-time process, such as traffic monitoring, web video and satellite map.
- Visual understanding: Vision and language is an interesting work related to scene text detection and recognition. Extracting text cues from scene and then recognizing text contribute to read and understand scene. Some methods focus on image classification and retrieval via text cues. For example, Mishra et al. [97] presented a query-driven search method for text-to-image retrieval. Karaoglu et al. [98] aim to handle fine-grained classification and business logo retrieval by combining the textual information and visual cues together.

6 Conclusion

This paper mainly described the basic situation of scene text detection and recognition in both image and video in three aspects: (1) text detection, (2) text recognition and (3) end-to-end text recognition system. We focused on the analysis of new methods and new ideas in the past 5 years. Compared with previous surveys, we paid more attention to deep learning-based methods. Moreover, we further introduced and analyzed some other related works such as script identification, text/non-text classification and text-to-image retrieval. Meanwhile, several widely used benchmark datasets and evaluation protocols were listed for reference. In addition, we also showed and compared the performances of state-of-the-art methods. Finally, we pointed out several potential trends in this field.

Acknowledgements This work was supported by the National Natural Science Foundation of China under Grants 61370039, and the Beijing Natural Science Foundation under Grant L172053.

References

1. Neumann, L., Matas, J.: A method for text localization and recognition in real-world images. In: Asian Conference on Computer Vision, pp. 770–783. Springer (2010)

2. Epshtein, B., Ofek, E., Wexler, Y.: Detecting text in natural scenes with stroke width transform. In: 2010 IEEE Conference on CVPR, pp. 2963–2970. IEEE (2010)
3. Yin, X.C., Yin, X., Huang, K., Hao, H.W.: Robust text detection in natural scene images. IEEE Trans. PAMI **36**(5), 970–983 (2014)
4. Neumann, L., Matas, J.: Real-time scene text localization and recognition. In: 2012 IEEE Conference on CVPR, pp. 3538–3545. IEEE (2012)
5. Cho, H., Sung, M., Jun, B.: Canny text detector: fast and robust scene text localization algorithm. In: CVPR, pp. 3566–3573 (2016)
6. Busta, M., Neumann, L., Matas, J.: Fastext: efficient unconstrained scene text detector. In: ICCV, pp. 1206–1214 (2015)
7. Zhong, Y., Zhang, H., Jain, A.K.: Automatic caption localization in compressed video. IEEE Trans. PAMI **22**(4), 385–392 (2000)
8. Hanif, S.M., Prevost, L., Negri, P.: A cascade detector for text detection in natural scene images. In: ICPR, pp. 1–4 (2008)
9. Hanif, S.M., Prevost, L.: Text detection and localization in complex scene images using constrained adaboost algorithm. In: ICDAR'09, pp. 1–5. IEEE (2009)
10. Zhang, Z., Shen, W., Yao, C., Bai, X.: Symmetry-based text line detection in natural scenes. In: CVPR, pp. 2558–2567 (2015)
11. Liang, G., Shivakumara, P., Lu, T., Tan, C.L.: A new wavelet-laplacian method for arbitrarily-oriented character segmentation in video text lines. In: ICDAR'15, pp. 926–930. IEEE (2015)
12. Huang, W., Qiao, Y., Tang, X.: Robust scene text detection with convolution neural network induced mser trees. In: ECCV, pp. 497–511. Springer (2014)
13. Zhong, Z., Sun, L., Huo, Q.: Improved localization accuracy by locnet for faster r-cnn based text detection. In: DICDAR'17, vol. 1, pp. 923–928. IEEE (2017)
14. Zhang, Z., Zhang, C., Shen, W., Yao, C., Liu, W., Bai, X.: Multi-oriented text detection with fully convolutional networks. In: CVPR, pp. 4159–4167 (2016)
15. Zhu, S., Zanibbi, R.: A text detection system for natural scenes with convolutional feature learning and cascaded classification. In: CVPR, pp. 625–632 (2016)
16. Qin, S., Manduchi, R.: Cascaded segmentation-detection networks for word-level text spotting. arXiv preprint arXiv:1704.00834 (2017)
17. Gupta, A., Vedaldi, A., Zisserman, A.: Synthetic data for text localisation in natural images. In: CVPR, pp. 2315–2324 (2016)
18. Tang, Y., Wu, X.: Scene text detection and segmentation based on cascaded convolution neural networks. IEEE Trans. Image Process. **26**(3), 1509–1520 (2017)
19. Wang, C., Yin, F., Liu, C.L.: Scene text detection with novel superpixel based character candidate extraction. In: ICDAR'17, vol. 1, pp. 929–934. IEEE (2017)
20. Turki, H., Halima, M.B., Alimi, A.M.: Text detection based on mser and cnn features. In: ICDAR'17, vol. 1, pp. 949–954. IEEE (2017)
21. Tian, S., Pan, Y., Huang, C., Lu, S., Yu, K., Lim Tan, C.: Text flow: a unified text detection system in natural scene images. In: ICCV, pp. 4651–4659 (2015)
22. Tian, Z., Huang, W., He, T., He, P., Qiao, Y.: Detecting text in natural image with connectionist text proposal network. In: ECCV, pp. 56–72. Springer (2016)
23. He, T., Huang, W., Qiao, Y., Yao, J.: Text-attentional convolutional neural network for scene text detection. IEEE Trans. Image Process. **25**(6), 2529–2541 (2016)
24. Fabrizio, J., Robert-Seidowsky, M., Dubuisson, S., Calarasanu, S., Boissel, R.: Textcatcher: a method to detect curved and challenging text in natural scenes. IJDAR **19**(2), 99–117 (2016)
25. Pei, W.Y., Yang, C., Kau, L.J., Yin, X.C.: Multi-orientation scene text detection with multi-information fusion. In: ICPR, pp. 657–662. IEEE (2016)
26. Yin, X.C., Pei, W.Y., Zhang, J., Hao, H.W.: Multi-orientation scene text detection with adaptive clustering. IEEE Trans. PAMI **37**(9), 1930–1937 (2015)
27. Kang, L., Li, Y., Doermann, D.: Orientation robust text line detection in natural images. In: CVPR, pp. 4034–4041 (2014)
28. Gomez, L., Karatzas, D.: Textproposals: a text-specific selective search algorithm for word spotting in the wild. Pattern Recognit. **70**, 60–74 (2017)
29. Zhou, X., Yao, C., Wen, H., Wang, Y., Zhou, S., He, W., Liang, J.: East: an efficient and accurate scene text detector. arXiv preprint arXiv:1704.03155 (2017)
30. Shi, B., Bai, X., Belongie, S.: Detecting oriented text in natural images by linking segments. In: CVPR, vol. 3 (2017)
31. Liu, Y., Jin, L.: Deep matching prior network: toward tighter multi-oriented text detection. In: CVPR, vol. 2, p. 8 (2017)
32. Sheshadri, K., Divvala, S.K.: Exemplar driven character recognition in the wild. In: BMVC, pp. 1–10 (2012)
33. Shi, C., Wang, C., Xiao, B., Zhang, Y., Gao, S., Zhang, Z.: Scene text recognition using part-based tree-structured character detection. In: CVPR, pp. 2961–2968. IEEE (2013)
34. Coates, A., Carpenter, B., Case, C., Satheesh, S., Suresh, B., Wang, T., Wu, D.J., Ng, A.Y.: Text detection and character recognition in scene images with unsupervised feature learning. In: ICDAR'11, pp. 440–445. IEEE (2011)
35. Yao, C., Bai, X., Shi, B., Liu, W.: Strokelets: a learned multi-scale representation for scene text recognition. In: CVPR, pp. 4042–4049 (2014)
36. Lee, C.Y., Bhardwaj, A., Di, W., Jagadeesh, V., Piramuthu, R.: Region-based discriminative feature pooling for scene text recognition. In: CVPR, pp. 4050–4057 (2014)
37. Lou, X., Kansky, K., Lehrach, W., Laan, C., Marthi, B., Phoenix, D., George, D.: Generative shape models: joint text recognition and segmentation with very little training data. In: NIPS, pp. 2793–2801 (2016)
38. Liang, G., Shivakumara, P., Lu, T., Tan, C.L.: Multi-spectral fusion based approach for arbitrarily oriented scene text detection in video images. IEEE Trans. Image Process. **24**(11), 4488–4501 (2015)
39. Elagouni, K., Garcia, C., Mamalet, F., Sébillot, P.: Combining multi-scale character recognition and linguistic knowledge for natural scene text OCR. In: 2012 10th IAPR International Workshop on Document Analysis Systems (DAS), pp. 120–124. IEEE (2012)
40. Phan, T.Q., Shivakumara, P., Tian, S., Tan, C.L.: Recognizing text with perspective distortion in natural scenes. In: ICCV, pp. 569–576. IEEE (2013)
41. Weinman, J.J., Butler, Z., Knoll, D., Feild, J.: Toward integrated scene text reading. IEEE Trans. PAMI **36**(2), 375–387 (2014)
42. Su, B., Lu, S.: Accurate scene text recognition based on recurrent neural network. In: ACCV, pp. 35–48. Springer (2014)
43. Ghosh, S.K., Valveny, E., Bagdanov, A.D.: Visual attention models for scene text recognition. arXiv preprint arXiv:1706.01487 (2017)
44. Shi, B., Wang, X., Lyu, P., Yao, C., Bai, X.: Robust scene text recognition with automatic rectification. In: CVPR, pp. 4168–4176 (2016)
45. Lee, C.Y., Osindero, S.: Recursive recurrent nets with attention modeling for OCR in the wild. In: CVPR, pp. 2231–2239 (2016)
46. He, P., Huang, W., Qiao, Y., Loy, C.C., Tang, X.: Reading scene text in deep convolutional sequences. AAAI **16**, 3501–3508 (2016)
47. Yang, X., He, D., Zhou, Z., Kifer, D., Giles, C.L.: Learning to read irregular text with attention mechanisms. In: IJCAI, pp. 3280–3286 (2017)

48. Shi, B., Bai, X., Yao, C.: An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. IEEE Trans. PAMI **39**(11), 2298–2304 (2017)
49. Neumann, L., Matas, J.: Scene text localization and recognition with oriented stroke detection. In: ICCV, pp. 97–104 (2013)
50. Jaderberg, M., Vedaldi, A., Zisserman, A.: Deep features for text spotting. In: ECCV, pp. 512–528. Springer (2014)
51. Jaderberg, M., Simonyan, K., Vedaldi, A., Zisserman, A.: Reading text in the wild with convolutional neural networks. IJCV **116**(1), 1–20 (2016)
52. Neumann, L., Matas, J.: Efficient scene text localization and recognition with local character refinement. In: ICDAR'15, pp. 746–750. IEEE (2015)
53. Neumann, L., Matas, J.: Real-time lexicon-free scene text localization and recognition. IEEE Trans. PAMI **38**(9), 1872–1885 (2016)
54. Yao, C., Bai, X., Liu, W.: A unified framework for multioriented text detection and recognition. IEEE Trans. Image Process. **23**(11), 4737–4749 (2014)
55. Liao, M., Shi, B., Bai, X., Wang, X., Liu, W.: Textboxes: a fast text detector with a single deep neural network. In: AAAI, pp. 4161–4167 (2017)
56. Ye, Q., Doermann, D.: Text detection and recognition in imagery: a survey. IEEE Trans. PAMI **37**(7), 1480–1500 (2015)
57. Zhu, Y., Yao, C., Bai, X.: Scene text detection and recognition: recent advances and future trends. Front. Comput. Sci. **10**(1), 19–36 (2016)
58. Yin, X.C., Zuo, Z.Y., Tian, S., Liu, C.L.: Text detection, tracking and recognition in video: a comprehensive survey. IEEE Trans. Image Process. **25**(6), 2752–2773 (2016)
59. Weinman, J.J.: Unified Detection and Recognition for Reading Text in Scene Images. University of Massachusetts Amherst, Amherst (2008)
60. Field, J.: Improving text recognition in images of natural scenes. PhD thesis, University of Massachusetts Amherst (2014)
61. Jaderberg, M.: Deep learning for text spotting. PhD thesis (2015)
62. Mishra, A.: Understanding Text in Scene Images. PhD thesis, International Institute of Information Technology Hyderabad (2016)
63. Bissacco, A., Cummins, M., Netzer, Y., Neven, H.: Photoocr: Reading text in uncontrolled conditions. In: ICCV, pp. 785–792. IEEE (2013)
64. Pan, Y.F., Hou, X., Liu, C.L.: Text localization in natural scene images based on conditional random field. In: ICDAR'09, pp. 6–10. IEEE (2009)
65. Pan, Y.F., Hou, X., Liu, C.L.: A hybrid approach to detect and localize texts in natural scene images. IEEE Trans. Image Process. **20**(3), 800–813 (2011)
66. Wang, Y., Shi, C., Xiao, B., Wang, C.: Mrf based text binarization in complex images using stroke feature. In: ICDAR'15, pp. 821–825. IEEE (2015)
67. Koo, H.I., Cho, N.I.: Text-line extraction in handwritten chinese documents based on an energy minimization framework. IEEE Trans. Image Process. **21**(3), 1169–1175 (2012)
68. Mishra, A., Alahari, K., Jawahar, C.: Top-down and bottom-up cues for scene text recognition. In: CVPR, pp. 2687–2694. IEEE (2012)
69. Sharma, N., Mandal, R., Sharma, R., Roy, P.P., Pal, U., Blumenstein, M.: Multi-lingual text recognition from video frames. In: ICDAR'15, pp. 951–955. IEEE (2015)
70. Canny, J.: A computational approach to edge detection. IEEE Trans. PAMI **8**, 679–698 (1986)
71. Fogel, I., Sagi, D.: Gabor filters as texture discriminator. Biol. Cybern. **61**(2), 103–113 (1989)
72. Mallat, S.G.: A theory for multiresolution signal decomposition: the wavelet representation. IEEE Trans. PAMI **11**(7), 674–693 (1989)
73. Van Loan, C.: Computational Frameworks for the Fast Fourier Transform. SIAM, Philadelphia (1992)
74. Jung, K., Kim, K.I., Jain, A.K.: Text information extraction in images and video: a survey. Pattern Recognit. **37**(5), 977–997 (2004)
75. Zuo, Z.Y., Tian, S., Pei, W.Y., Yin, X.C.: Multi-strategy tracking based text detection in scene videos. In: ICDAR'15, pp. 66–70. IEEE (2015)
76. Tian, S., Yin, X.C., Su, Y., Hao, H.W.: A unified framework for tracking based text detection and recognition from web videos. IEEE Trans. PAMI **40**(3), 542–554 (2018)
77. Shivakumara, P., Phan, T.Q., Tan, C.L.: A laplacian approach to multi-oriented text detection in video. IEEE Trans. PAMI **33**(2), 412–419 (2011)
78. Yousfi, S., Berrani, S.A., Garcia, C.: Deep learning and recurrent connectionist-based approaches for arabic text recognition in videos. In: ICDAR'15, pp. 1026–1030. IEEE (2015)
79. Yao, C., Bai, X., Liu, W., Ma, Y., Tu, Z.: Detecting texts of arbitrary orientations in natural images. In: CVPR, pp. 1083–1090. IEEE (2012)
80. Nicolaou, A., Bagdanov, A.D., Gómez, L., Karatzas, D.: Visual script and language identification. In: 2016 12th IAPR Workshop on Document Analysis Systems (DAS), pp. 393–398. IEEE (2016)
81. Shi, B., Bai, X., Yao, C.: Script identification in the wild via discriminative convolutional neural network. Pattern Recognit. **52**, 448–458 (2016)
82. Gomez, L., Nicolaou, A., Karatzas, D.: Improving patch-based scene text script identification with ensembles of conjoined networks. Pattern Recognit. **67**, 85–96 (2017)
83. Sharma, N., Mandal, R., Sharma, R., Pal, U., Blumenstein, M.: ICDAR 2015 competition on video script identification (cvsi 2015). In: ICDAR'15, pp. 1196–1200. IEEE (2015)
84. Delaye, A., Liu, C.L.: Contextual text/non-text stroke classification in online handwritten notes with conditional random fields. Pattern Recognit. **47**(3), 959–968 (2014)
85. Van Phan, T., Nakagawa, M.: Text/non-text classification in online handwritten documents with recurrent neural networks. In: ICFHR, pp. 23–28. IEEE (2014)
86. Sharma, N., Shivakumara, P., Pal, U., Blumenstein, M., Tan, C.L.: Piece-wise linearity based method for text frame classification in video. Pattern Recognit. **48**(3), 862–881 (2015)
87. Bai, X., Shi, B., Zhang, C., Cai, X., Qi, L.: Text/non-text image classification in the wild with convolutional neural networks. Pattern Recognit. **66**, 437–446 (2017)
88. Otsu, N.: A threshold selection method from gray-level histograms. IEEE Trans. Syst. Man Cybern. **9**(1), 62–66 (1979)
89. Sauvola, J., Pietikäinen, M.: Adaptive document image binarization. Pattern Recognit. **33**(2), 225–236 (2000)
90. Yi, C., Tian, Y.: Localizing text in scene images by boundary clustering, stroke segmentation, and string fragment classification. IEEE Trans. Image Process. **21**(9), 4256–4268 (2012)
91. Howe, N.R.: Document binarization with automatic parameter tuning. IJDAR **16**(3), 247–258 (2013)
92. Zhang, Z., Wang, W.: A novel approach for binarization of overlay text. In: 2013 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 4259–4264. IEEE (2013)
93. Tensmeyer, C., Martinez, T.: Document image binarization with fully convolutional neural networks. arXiv preprint arXiv:1708.03276 (2017)
94. Peng, X., Cao, H., Natarajan, P.: Using convolutional encoder–decoder for document image binarization. In: ICDAR'17, vol. 1, pp. 708–713. IEEE (2017)

95. Meng, G., Yuan, K., Wu, Y., Xiang, S., Pan, C.: Deep networks for degraded document image binarization through pyramid reconstruction. In: ICDAR'17, vol. 1, pp. 727–732. IEEE (2017)
96. Ha, J.W., Lee, B.J., Zhang, B.T.: Text-to-image retrieval based on incremental association via multimodal hypernetworks. In: 2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 3245–3250. IEEE (2012)
97. Mishra, A., Alahari, K., Jawahar, C.: Image retrieval using textual cues. In: ICCV, pp. 3040–3047. IEEE (2013)
98. Karaoglu, S., Tao, R., Gevers, T., Smeulders, A.W.M.: Words matter: scene text for image classification and retrieval. IEEE Trans. Multimed. **19**(5), 1063–1076 (2017)
99. Rong, X., Yi, C., Tian, Y.: Unambiguous text localization and retrieval for cluttered scenes. In: CVPR, pp. 3279–3287. IEEE (2017)
100. Lucas, S.M., Panaretos, A., Sosa, L., Tang, A., Wong, S., Young, R.: ICDAR 2003 robust reading competitions. In: ICDAR'03, pp. 682–687. IEEE (2003)
101. Lucas, SM.: ICDAR 2005 text locating competition results. In: ICDAR'05, pp. 80–84. IEEE (2005)
102. Shahab, A., Shafait, F., Dengel, A.: ICDAR 2011 robust reading competition challenge 2: reading text in scene images. In: ICDAR'11, pp. 1491–1496. IEEE (2011)
103. Karatzas, D., Shafait, F., Uchida, S., Iwamura, M., i Bigorda, L.G., Mestre, S.R., Mas, J., Mota, D.F., Almazan, J.A., de las Heras, L.P.: ICDAR 2013 robust reading competition. In: ICDAR'13, pp. 1484–1493. IEEE (2013)
104. Karatzas, D., Gomez-Bigorda, L., Nicolaou, A., Ghosh, S., Bagdanov, A., Iwamura, M., Matas, J., Neumann, L., Chandrasekhar, V.R., Lu, S., et al.: ICDAR 2015 competition on robust reading. In: ICDAR'15, pp. 1156–1160. IEEE (2015)
105. Veit, A., Matera, T., Neumann, L., Matas, J., Belongie, S.: Coco-text: Dataset and benchmark for text detection and recognition in natural images. arXiv preprint arXiv:1601.07140 (2016)
106. Mishra, A., Alahari, K., Jawahar, C.: Scene text recognition using higher order language priors. In: BMVC, BMVA (2012)
107. Campos, T.E.D., Babu, B.R., Varma, A.M.: Character Recognition in Natural Images. Chapman & Hall, Boca Raton (2009)
108. SeongHun, L., Min Su, C., Kyomin, J., Jin Hyung, K.: Scene text extraction with edge constraint and text collinearity. In: 2010 20th International Conference on Pattern Recognition, pp. 3983–3986. IEEE (2010)
109. Yi, C., Tian, Y.: Text string detection from natural scenes by structure-based partition and grouping. IEEE Trans. Image Process. **20**(9), 2594–2605 (2011)
110. Ch'ng, C.K., Chan, C.S.: Total-text: a comprehensive dataset for scene text detection and recognition. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), vol. 1, pp. 935–942. IEEE (2017)
111. Shi, B., Yao, C., Liao, M., Yang, M., Xu, P., Cui, L., Belongie, S., Lu, S., Bai, X.: ICDAR 2017 competition on reading chinese text in the wild (rctw-17). arXiv preprint arXiv:1708.09585 (2017)
112. Risnumawan, A., Shivakumara, P., Chan, C.S., Tan, C.L.: A robust arbitrary text detection system for natural scene images. Expert Syst. Appl. **41**(18), 8027–8048 (2014)
113. Wolf, C., Jolion, J.M.: Object count/area graphs for the evaluation of object detection and segmentation algorithms. IJDAR **8**(4), 280–296 (2006)
114. Cheng, Z., Bai, F., Xu, Y., Zheng, G., Pu, S., Zhou, S.L Focusing attention: towards accurate text recognition in natural images. In: ICCV, pp. 5086–5094. IEEE (2017)
115. Jaderberg, M., Simonyan, K., Vedaldi, A., Zisserman, A.: Deep structured output learning for unconstrained text recognition. In: ICLR (2015)
116. Alsharif, O., Pineau, J.: End-to-end text recognition with hybrid hmm maxout models. arXiv preprint arXiv:1310.1811 (2013)
117. Wang, K., Babenko, B., Belongie, S.: End-to-end scene text recognition. In: ICCV, pp. 1457–1464. IEEE (2011)
118. Li, H., Wang, P., Shen, C.: Towards end-to-end text spotting with convolutional recurrent neural networks. In: Proc. ICCV, pp. 5238–5246 (2017)
119. Hu, H., Zhang, C., Luo, Y., Wang, Y., Han, J., Ding, E.: Wordsup: exploiting word annotations for character based text detection. In: ICCV (2017)
120. He, W., Zhang, X.Y., Yin, F., Liu, C.L.: Deep direct regression for multi-oriented scene text detection. arXiv preprint arXiv:1703.08289 (2017)
121. He, P., Huang, W., He, T., Zhu, Q., Qiao, Y., Li, X.: Single shot text detector with regional attention. In: ICCV (2017)
122. Busta, M., Neumann, L., Matas, J.: Deep textspotter: an end-to-end trainable scene text localization and recognition framework. In: ICCV, pp. 22–29 (2017)
123. Wu, Y., Natarajan, P.: Self-organized text detection with minimal post-processing via border learning. In: CVPR, pp. 5000–5009 (2017)
124. Gordo, A.: Supervised mid-level features for word image representation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2956–2964 (2015)
125. Almazan, J., Gordo, A., Fornes, A., Valveny, E.: Word spotting and recognition with embedded attributes. IEEE Trans. PAMI **36**(12), 2552–2566 (2014)