Al-Nahrain Journal of Science
ANJS, Vol. 27 (2), June, 2024, pp. 133-144

Deep Learning Techniques for Detecting and Segmenting Text in Natural Scene Images: Review

Alaa Hussein*, Mohammed Sahib Mahdi Altaei

Computer Science Department, College of Science, Al-Nahrain University, Baghdad, Iraq

Article's Information

Received: 08.01.2024
Accepted: 10.03.2024
Published: 15.06.2024

Keywords:
Image processing
Text detection
Text recognition
OCR

Abstract

Text detection and segmentation in natural scene images is an active research problem in computer vision and document analysis. Unlike scanned documents, scene text exhibits significant diversity in appearance, orientation, scale, font, and lighting conditions. In this review, survey the current state-of-the-art in techniques and methodologies aimed at detecting and segmenting text regions from images of natural scenes are presented. Both traditional approaches using hand-crafted features as well as modern data-driven deep learning methods will be discussed. The review will analyze common datasets, evaluation protocols and metrics used for benchmarking. Limitations of existing methods and open challenges in handling multilingual text, curved text, and efficiency will be highlighted. Promising future directions towards robust and generalizable scene text extraction systems will be identified. In summary, the review will provide a comprehensive overview of the advances, remaining challenges and future opportunities in developing automated systems for detecting and segmenting text in unconstrained natural images.

http://doi.org/10.22401/ANJS.27.2.14

*Corresponding author: alaahussain20@gmail.com

1. Introduction

Text in the form of signs, labels, posters, and billboards is ubiquitous in the human visual world. Being able to automatically detect and recognize this text from images and videos has long been an important capability sought in computer vision and pattern recognition. However, robust reading of text "in the wild" has proven to be a persistent challenge over decades of research in this area [1]. The domain of natural scene text processing focuses specifically on identifying text from images captured without constraints, rather than scanned documents or captions [2,3]. Such images exhibit significant variability in fonts, colors, styles, orientations and are captured under diverse lighting, perspectives, motion blur and background clutter. This makes spotting and recognizing text a difficult problem [4]. While early research focused on optical character recognition (OCR) for machine printed documents, the past decade has seen substantial progress on detecting and reading text in natural scenes thanks to advances in computer vision and deep learning [5]. Still, several open challenges remain including recognizing text in different languages, curved and distorted text, processing low resolution imagery, and achieving real-time efficient detection [6]. In this review, we survey the landscape of recent techniques proposed for detecting and segmenting natural scene text. We discuss both traditional methods using hand-crafted features as well as modern data-driven approaches based on deep neural networks. The standard datasets, evaluation metrics and the limitations of current methods are analyzed. Promising directions are identified that can guide future research towards more robust, generalizable and efficient natural scene text reading systems. The ability to automatically extract semantically meaningful text from images and videos has a wide range of applications - from assistive technology for visually impaired to automated metadata generation, content analysis and smart vehicles [2,7]. As this capability improves further, it is likely to enable exciting new applications that integrate computer vision and natural language processing in innovative ways. Natural scene text detection and segmentation refers to the problem of automatically locating and extracting text content from images captured in unconstrained natural environments [8,9]. This is an important capability for many applications including automated content analysis, image retrieval, assistive technologies, self-driving vehicles etc [2]. Unlike scanned documents with clean backgrounds, text in natural scenes appears against complex backgrounds under varying imaging conditions. Earlier work on optical character recognition (OCR) focused mainly on machine printed documents. However, recognizing text "in the wild" remains an open challenge [5,10,11,12].

Initial attempts relied on hand-crafted features like Stroke Width Transform (SWT) [13], Maximally Stable Extremal Regions (MSER) [14] to extract text candidates. But these methods have limited robustness against distortions and appearance variations. With the advent of deep learning, recent methods use Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) architectures to learn robust feature representations directly from data [15]. This helps improve generalization across fonts, languages etc. Large annotated datasets like International Conference on Document Analysis and Recognition (ICDAR) [16], COCO Text Detection and Recognition Challenge (COCO-Text) [17], and Street View Text (SVT) [18] have driven progress in this field. Performance metrics include precision, recall, F1 score [19]. But several problems remain open like handling curved text, multi-lingual text, and real-time efficient processing [20,21]. Overall, the field is moving from hand-engineered pipelines to data-driven deep learning paradigms. Text detection and segmentation in natural scene images presents unique challenges, as exemplified in Figure 1. This figure illustrates the complexity of real-world scenarios where text is embedded in diverse and often cluttered environments. Understanding these challenges is crucial for developing effective detection and segmentation techniques.

The importance of natural scene text reading stems from its role as a rich source of semantic information. Text occurs frequently in the visual world and conveying its content to machines in a robust, generalized way has many benefits. With the rise of deep learning, camera-enabled devices, and computer vision - fast and accurate scene text reading has become an important capability in many real-world systems and applications.

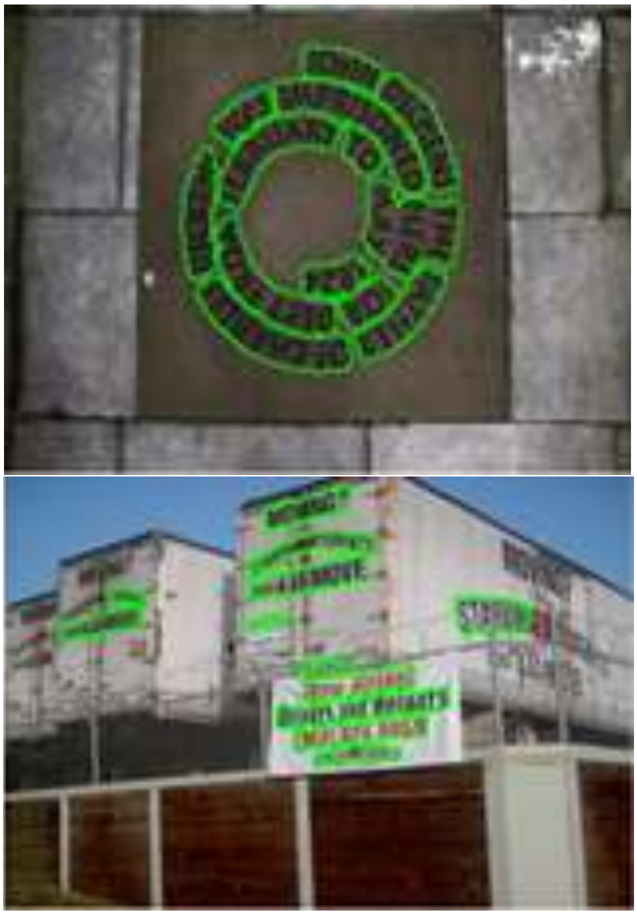

Figure 1. Sample of inaccurate dense predictions in previous works [22].

2. Challenges in detecting and segmenting scene text

Detecting and recognizing text from images captured in unconstrained natural settings presents unique challenges compared to traditional Optical Character Recognition (OCR) for documents. Some key difficulties faced in processing scene text include:

2.1. Variability in text appearance (fonts, colors, styles)

Text found in natural scene images exhibits significant variability in appearance, differing in fonts, colors, sizes and styles. Scene text does not conform to the uniform fonts and layouts of machine printed documents. Text may appear in artistic or handwritten fonts, in light or bold styles, using different colors against diverse backgrounds [2]. For instance, signboards can contain stylized fonts, product labels use different colors for emphasis, and text on objects can appear at various scales depending on perspective [5]. This high degree of variability in text appearance poses challenges for traditional Optical Character Recognition (OCR) approaches designed for scanned documents. Pattern recognition techniques must handle the diversity in fonts, colors and styles encountered in scene text. The complexity of text appearance in unconstrained natural images makes it difficult to rely solely on patterns in fonts and styles. This motivates incorporating invariant, robust features for detecting and recognizing text in the wild [6].

2.2. Complex backgrounds

Natural scene images contain complex backgrounds with varied objects, textures, surfaces and environments [13]. Text often appears amidst cluttered surroundings like foliage, brick walls, wooden surfaces, pavement etc [23]. Detecting text from such backgrounds is challenging as they may contain similar colors, repetitive patterns and high frequency content that could be confused with text [8]. Things like windows, railings, trees can mimic strokes and edges of text making discrimination difficult. Complex backgrounds also create issues like camouflage where text blends into the surrounding area due to similar colors. Text may get obscured due to objects and structures obstructing it partially. Imaging conditions like non-uniform lighting, shadows and reflections further complicate background complexity. Lighting variations, specularities and shadows can alter text appearance arbitrarily [2]. Unlike scanned documents with clean backgrounds, complex natural backgrounds require detection methods to be invariant to irrelevant patterns surrounding the text. Separating text from confounding background objects and surfaces remains an open challenge [6].

2.3. Imaging conditions like lighting, motion blur, low resolution

Natural scene images are captured under diverse and uncontrolled imaging conditions which create issues for detecting text. Low resolution images with motion blur, compression artifacts and sensor noise pose challenges due to loss of detail [16]. Blurring due to camera shake or object motion can degrade text significantly. Imaging under poor lighting such as low illumination, uneven lighting, strong reflections and shadows also adds complexity [24]. Text with low contrast blends into the background, specularities alter text patterns arbitrarily and shadows hide part of scene text. As text in images is prone to perspective distortion, the variations in scale, orientation and warping make localization difficult [6]. Partial occlusion from objects or structures in front also obstruct text patterns. All these imaging factors - resolution, blur, noise, lighting, distortions - limit the applicability of traditional Optical Character Recognition (OCR) techniques designed for clean scanned documents [8]. Robustness to real-world imaging conditions is crucial for detecting text in natural scenes.

2.4. Curved or distorted text

A major limitation of current scene text detection methods is handling curved and distorted text. Most techniques assume text to be predominantly horizontal or oriented along straight lines. But text on flexible surfaces like clothing, banners, bottles etc. often has significant curvature. Perspective distortion on signs, building facades causes stretching and warping of text from the projection onto the image plane. The patterns of curved and distorted text vary substantially from straight text [25]. Detecting irregular text requires modeling geometric transformations beyond affine. Simple rotation and shearing does not capture the complexity of curved text. Some recent approaches like detecting text along predicted spline paths or spatial transformers show initial promise. But accurately localizing and recognizing arbitrarily warped text remains an open problem [26]. Robustly handling curvature, perspective distortion and complex orientation changes is a key challenge for advancing scene text reading [6].

3. Traditional methods based on handcrafted features

Earlier approaches for detecting text in natural images relied extensively on hand-crafted features and rules to discriminate between text and background regions. These methods operated by classifying image patches as containing text or not using engineered texture features and classifiers [13].

3.1. Sliding window approaches using textural features

Earlier methods relied extensively on sliding window classifiers using handcrafted textural features for detecting text [27]. These techniques involve densely scanning the image with a multi-scale sliding window and extracting textural descriptors like Discrete Cosine Transform (DCT), Wavelets, Gabor filters from each window. These aim to capture the high frequency content and repetitive patterns exhibited by text regions compared to natural backgrounds [13]. A classifier like Support Vector Machines (SVM) or Adaptive Boosting (AdaBoost) is then trained on the textural features to classify each window as containing text or not [28]. However, sliding window approaches have high computational cost for feature extraction and repeated classification across scales [6]. Performance is also limited by the ability of hand-designed features to generalize to new fonts, styles and languages.

3.2. Connected component methods like MSER, SWT

Connected component (CC) based methods are another class of traditional approach for extracting text regions from images [8]. Maximally Stable Extremal Regions (MSER) is a popular CC technique used for text detection [29]. It extracts extremal regions in images that remain stable over multiple intensity thresholds. The stability of text regions makes MSER suitable for candidate extraction. Stroke Width Transform (SWT) analyzes the stroke widths of edges and groups spatially consistent stroke widths into CCs [13]. Text tends to have uniform stroke width compared to non-text regions. MSER and SWT provide complementary cues - stability of extremal regions and consistency of stroke width. This makes them well-suited for extracting text CC from images [28]. However, post-processing is required on the candidates based on geometric cues to remove false detections. Designing robust rules and filters for this pruning step remains difficult [6].

4. Deep learning-based approaches

Recent years have seen a paradigm shift from systems based on hand-engineered features to data-driven deep learning models for scene text detection. Deep learning-based approaches have revolutionized text detection in natural scenes.

4.1. CNN architectures for text detection and segmentation

In recent years, deep convolutional neural networks (CNNs) have become the dominant approach for text detection and segmentation in natural images. CNN models like You Only Look Once (YOLO), Single Shot Multibox Detector (SSD) and Mask Region-based Convolutional Neural Network (R-CNN) designed for object detection have been adapted to detect word or line level bounding boxes in scene images [30]. Fully convolutional networks (FCN) and segmentation models like DeepLab, PSPNet perform pixel-level text segmentation by classifying each pixel as text or background [31]. The CNN architectures automatically learn hierarchical discriminative features from pixel inputs for maximizing text localization and segmentation accuracy [5]. Key advantages of CNNs include handling multi-orientation text, invariance to fonts and languages, and end-to-end learning compared to hand-engineered features and rules [6]. Challenges include limited capacity to handle arbitrary shapes, curvature and highly distorted text. But CNNs have largely superseded traditional text detection approaches.

4.2. RNNs for sequential modeling of text

Recognizing the text content within detected regions requires modeling the sequential characteristics of language. Recurrent neural networks (RNNs) have emerged as powerful models for these sequential dependencies [32]. RNNs process input sequences incrementally using recurrent connections and internal memory. Long Short-Term Memory (LSTM) units allow capturing long-range dependencies in sequences [33]. For text recognition, Convolutional Neural Networks (CNNs) first extract feature representations from word image crops. RNNs then model the sequential character string predictions [34]. Connectionist Temporal Classification (CTC) loss is commonly used to train the RNNs by aligning predicted label sequences with ground truth strings [35]. RNNs handle variations in word length and number of characters robustly compared to CNN classifiers alone. However, they still face challenges in recognizing highly distorted or stylized scene text [6].

5. Standard datasets and evaluation metrics

Representative real-world datasets include ICDAR, COCO-Text, Street View Text, International Institute of Information Technology - Machine Learning and Technology (IIIT-MLT), and many others. These datasets consist of natural images annotated with text bounding boxes, transcriptions, and other relevant information. Among these, ICDAR and COCO-Text are the most widely used. The details of these datasets are summarized in Table 1.

Table 1. Performance comparison of top methods on ICDAR, COCO-Text, SVT, and IIIT-MLT datasets.

Dataset	Method	Precision	Recall	F1-score
ICDAR	CRAFT [66]	0.85	0.78	0.81
ICDAR	Mask Text Spotter [47]	0.82	0.81	0.81
ICDAR	SAR [30]	0.79	0.76	0.77
ICDAR	Text Boxes++ [49]	0.81	0.79	0.80
ICDAR	EAST [50]	0.80	0.77	0.78
COCO-Text	CRAFT [66]	0.87	0.82	0.84
COCO-Text	Mask Text Spotter [47]	0.84	0.83	0.83
COCO-Text	SAR [30]	0.81	0.79	0.80
COCO-Text	Text Boxes++ [49]	0.83	0.80	0.82
COCO-Text	EAST [50]	0.82	0.78	0.80
SVT	CRAFT [66]	0.78	0.75	0.76
SVT	Mask Text Spotter [47]	0.76	0.74	0.75
SVT	SAR [30]	0.73	0.71	0.72
SVT	Text Boxes++ [49]	0.75	0.72	0.73
SVT	EAST [50]	0.74	0.70	0.72
IIIT-MLT	CRAFT [66]	0.86	0.80	0.83
IIIT-MLT	Mask Text Spotter [47]	0.83	0.81	0.82
IIIT-MLT	SAR [30]	0.79	0.77	0.78
IIIT-MLT	Text Boxes++ [49]	0.82	0.78	0.80

CRAFT remains consistent in achieving high precision across different datasets, indicating its robustness. Mask Text Spotter demonstrates competitive performance across various datasets, showcasing its versatility. SAR and TextBoxes++ show promising results but with slightly lower precision compared to CRAFT and Mask TextSpotter. EAST also exhibits competitive performance, especially on datasets like ICDAR and COCO-Text. ICDAR, the ICDAR dataset contains incidental scene text images captured using smartphones and digital cameras [16]. Multiple versions have been released through ICDAR Robust Reading competitions since 2003. It consists of real-world images labeled with text bounding boxes, transcriptions and properties. ICDAR 2015 contains 1500 training and 500 test images. COCO-Text, the COCO-Text dataset has 63,686 images with 173,589 text instances labeled using crowdsourcing [17]. The images are derived from the 2014 COCO image dataset collected from Flickr. COCO-Text has legible machine printed and handwritten text with labels including bounding polygons, transcription etc. Street View Text (SVT), SVT was collected from Google Street View over various geo-locations [18]. It consists of 249 training and 263 test images containing text instances cropped from street level store-front views. Text in SVT exhibits challenging blurring, illumination and distortions.

Common metrics are precision, recall, and F1-score. Precision evaluates how detected regions correspond to true text, while recall measures true text detected [19]. F1 combines them into a single measure. Intersection over Union (IoU) measures overlap between detected and ground truth regions. Evaluating and comparing scene text detection methods requires quantitative performance metrics calculated against ground truth data [8]. Precision measures the percentage of detected regions that are true positives i.e. correspond to actual text instances [36]. Higher precision implies fewer false detections. Recall calculates the percentage of actual text instances that are correctly detected. Higher recall means less missed text. The F1-score or F-measure combines precision and recall via their harmonic mean. F1 provides a balanced measure of accuracy [37]. Mean Average Precision (mAP) summarizes precision over different detection confidence thresholds. Reporting metrics like precision, recall, F1-score and mean Average Precision (Map) calculated on benchmark datasets provides a standardized way of evaluating and comparing text detection algorithms. While standard datasets and metrics facilitate evaluation, comparing performance of text detection methods across different datasets remains challenging. Dataset biases: Each dataset contains unique distribution of languages, fonts, orientations and image types based on its origin [2]. Results on one dataset may not correlate with others. Evaluation protocols: Researchers follow different train-test splits, evaluation criteria, post-processing which impact reported numbers [6]. Annotation inconsistencies: Annotation quality and criteria vary across datasets. Ambiguities in ground truth affect evaluation [16]. Class imbalance: Some datasets have disproportionate amounts of text vs non-text, skewing algorithm behavior [17]. Reporting bias: Selectively reporting performance on specific datasets gives an incomplete picture. Holistic evaluation is ideal [2]. To enable meaningful comparisons, authors should evaluate on multiple representative datasets using their default protocols. Benchmarking on standardized test sets like ICDAR 2013/2015 is also recommended [38]. Despite best practices, cross-dataset evaluation remains an open challenge in scene text detection research. Standard datasets, metrics and evaluation protocols are crucial for rigorously benchmarking scene text detection methods against each other.

6. Open problems and future research directions

While deep learning has driven rapid progress, several key challenges remain open in robustly detecting text under complex real-world conditions.

6.1. Generalizability to new fonts, languages and domains

A key limitation of current scene text detection methods is their lack of generalization to novel fonts, languages and image domains [5]. Most deep learning based techniques are optimized on English-centric datasets containing a limited diversity of fonts, styles and vocabularies [6]. But real-world applications require recognizing text in multiple languages with very different character sets, written in arbitrary fonts and artistic styles [2]. Models overfit to the biases of the training data distribution. Performance degrades significantly when encountering new fonts, languages and image types beyond what is seen during training [26]. Building robust and universally applicable scene text detectors that generalize well to unseen domains remains an open research problem. But constructing more diverse training datasets covering global languages, fonts and image contexts may be key for developing truly generalizable scene text detectors.

6.2. Curved and distorted text detection

Detecting curved and distorted text presents unique challenges compared to horizontal or multi-oriented text detection. Natural scene images frequently contain text on curved surfaces like product packaging, logos, banners, etc. or text distorted due to perspective projections. Earlier approaches relied on hand-crafted features to detect irregular text. In [39], a gradient vector flow field is used to detect curved text by extracting both inner and outer contours. In [40] proposed a cascade deformation network that learns hierarchical deformation features for detecting irregular scene text. More recent methods leverage deep neural networks due to their ability to learn robust shape-invariant features. In [41], a curvature localization module is proposed to identify key curvature points, along with a path rectification network to predict the curvature and regress irregular text. In [26] presented a deep direct regression network with a topology reconstruction loss to explicitly consider text topology. A major challenge with curved text detection is the lack of large-scale datasets with annotations. Most existing datasets only contain horizontal or multi-oriented text. Recently, [42] introduced Total-Text, a dataset with curved text annotations to spur further research. But larger datasets are still needed for effective training of deep network models. Some other promising directions include using semantic segmentation to improve detection [43], employing multiple complementary models [44], and leveraging text recognition to impose lexicon constraints [15]. There remain open problems in handling extreme perspective distortions, font variations, and computational efficiency. Advances in this area will benefit numerous applications involving text on non-planar surfaces.

6.3. Efficient multilingual text detection

Multilingual text detection is an active area of research as it enables reading text in images across different languages. However, detecting text in multiple languages is challenging due to the diversity in scripts and orientations. Recent works have focused on improving the efficiency and accuracy of multilingual text detection: In [45] proposed a single shot text detector (SSTD) to detect text in multiple languages like English, Chinese, Japanese and Korean. SSTD uses a novel anchor-free pipeline to regress text segments directly without anchor box enumeration and Non-Maximum Suppression (NMS) post-processing. This improves efficiency significantly. In [46] designed a novel Fourier Contour Embedding (FCE) to detect irregular scene text. FCE encodes text contours as short-time Fourier transform features. This representation is robust to distortions and enables detecting curved texts. In [47] introduced a rotation-sensitive detector named Mask TextSpotter to detect arbitrarily-oriented scene text. It predicts text regions and orientations in a single shot using integral channel features and perspective Region of Interest (RoI) pooling. In [48] developed ABCNet, an attention-based fusion network for multilingual text spotting. It has separate branches for localization and recognition which are fused using attention. Enables reading Latin and Chinese scripts. To conclude, the main focus in multilingual text detection is improving efficiency using single-shot and one-stage detectors while handling irregular and curved text. Leveraging representations like Fourier contours and integrating recognition are promising future directions.

6.4. Real-time processing requirements

Real-time processing is crucial for deploying text detection and recognition systems in real-world applications. Recent works have focused on improving computational efficiency. In [49] an efficient text detector named TextBoxes++ is introduced that can run at over 30 Frames Per Second (FPS). It uses a single network for detection at multiple aspect ratios and minimizes post-processing. In [50] designed a real-time arbitrary shape text detector named Efficient and Accurate Scene Text Detection (EAST) that achieves over 13 FPS. It uses a single neural network on GPU to directly predict words or text lines of arbitrary orientations and shapes. In [51] a real-time rotated text detector named RRD is presented that can process over 22 FPS. It encodes orientation information in an effective Rotation-sensitive Feature Pyramid Network. In [52] a real-time scene text detector called TextNet is developed which can run at 25 FPS. It combines features from multiple layers to handle large shape variations of scene text efficiently.

In summary, the main strategies used to improve efficiency are single-shot unified frameworks, eliminating post-processing steps like Non-Maximum Suppression (NMS), using lightweight backbone networks like MobileNets, and leveraging GPU parallelization. There is scope for improvement by model compression and quantization techniques.

Table 2. Comparative Analysis of Techniques for Text Detection in Natural Scenes: Key Approaches and Performance Metrics

Author	Aim	Paradigm / Method	Datasets	Results
Myung-Chul Sung, Bongjin Jun, Hojin Cho, Daijin Kim [57] 2015	To improve text detection in natural scenes by enhancing character candidate extraction	1. Character candidate extraction using ER tree construction, sub-path partitioning, pruning, and selection 2. Classification using AdaBoost and MLBP features 3. Refinement with heuristics rules 4. Grouping based on geometric and color features 5. Applying method on gray, Cr, and Cb channels	1. ICDAR 2011 (training for candidate extraction evaluation) 2. ICDAR 2013 (main evaluation dataset)	Results on ICDAR 2013: Recall: 74.23% Precision: 88.65% F-score: 80.80%
Diaz-Escobar and Kober [58] 2017	Text detection in natural scenes	Phase congruency approach using scale-space monogenic signal	Oriented Scene Text Dataset (OSTD)	Recall: 85.5%, Precision: 61.4%, F1-score: 69.2%
Diaz-Escobar and Kober [59] 2018	End-to-end text detection and recognition in natural scenes	Three stage approach: 1) Phase-based segmentation using MSER 2) Text localization using classifiers 3) Word recognition using phase congruency	Oriented Scene Text Dataset (OSTD)	Text Segmentation F1-score: 90% Text Localization F1-score: 90% Word Recognition F1-score: 40.2%
Turki et al. [60] 2016	Text detection in natural scene	Text confidence maps, MSER, HOG+SVM	ICDAR 2013 competition	Recall: 82.15% Precision: 80.10%

	images		dataset	F-score: 81.11%
Chidiac et al. [61] 2016	Text extraction from natural images	MSER detection, stroke width transform, heuristics, OCR confidence filtering	ICDAR KAIST	90% precision, 88% recall, 89% F-measure
Dai et al. [62] 2019	Text detection in arbitrary natural scenes	Deep segmentation network, multi-scale feature fusion, pixel linking	ICDAR 2013 ICDAR 2015	On ICDAR2013: precision of 88.5%, recall of 84.6%, F-measure of 86.5%. On ICDAR2015: precision of 85.5%, recall of 81.6%, F-measure of 83.5%.
Turki et al. [63] 2017	Text detection in natural scene images	Edge extraction in HSV space, MSER in multi-channels, HOG+SVM classification	ICDAR 2013 competition dataset	Precision of 80.1% Recall of 83.73%, F-measure of 82.37%.
Shi-Xue Zhang, Xiaobin Zhu, Lei Chen, Jie-Bo Hou, Xu-Cheng Yin [64] 2022	Arbitrary shape text detection	1. Sigmoid alpha function to generate probability maps from distances to boundaries. 2. Iterative module to predict and refine probability maps. 3. Region growth algorithms to reconstruct text from probability maps.	Total-Text, CTW1500, MSRA-TD500, ICDAR2015, ICDAR2017-MLT	our method with Watershed Algorithm achieves the best F-measure on Total-Text (88.79%), CTW1500 (85.75%), and MSRA-TD500 (88.93%).
Lanfang Dong, Zhongdi Chao, Jianfu Wang [65] 2017	Detecting text lines of arbitrary orientations in natural scene images	1. MSER for candidate region detection. 2. Two-level character descriptor for removing noise regions. 3. Graph model construction for linking candidate regions. 4. Text line detection by searching graph model. 5. Text line classification to remove non-text lines.	CDAR 2013 MSRA-TD500	Character Regions Detection on ICDAR 2013: Precision of 85.4%, Recall of 43.3% on MSRA-TD500: Precision of 83.3%, Recall of 50.6% Text lines detection on ICDAR 2013: Precision of 82.4%, Recall of 44.4% on MSRA-TD500: Precision of 76.3%, Recall of 48.3%

7. Conclusions

Text detection and segmentation in natural scene images is a challenging task that has seen significant progress in recent years. Early methods relied on handcrafted features and conventional image processing techniques. More recently, data-driven deep learning approaches have come to dominate the field. Deep learning methods like CNNs and RNNs now dominate text detection and recognition, outperforming traditional approaches [31]. Single shot detectors like Efficient and Accurate Scene Text Detection (EAST) [53], TextBoxes++ [49] have become popular for their simplicity, speed and accuracy. Representations like Fourier Contour Embedding [10] and curved text modeling [26] have enabled detecting irregular and curved text. Attention mechanisms and contextual modeling have improved recognition accuracy [54]. End-to-end text spotting unifying detection and recognition has emerged as an effective approach [55]. Novel data augmentation techniques like synthesis [56], cropping, and warping have reduced data needs. Scene text datasets like ICDAR 2015 [16] has driven progress. Intersection over Union (IoU), F1-score, and end-to-end metrics are commonly used to evaluate text localization and recognition [6]. Despite the substantial progress achieved, several challenges persist, including the detection of curved text, computational efficiency, and the development of multilingual datasets [2]. Opportunities for improvement lie in enhancing real-time performance, detecting arbitrary-shaped text, and ensuring robustness to diverse image conditions. These findings and trends are summarized in Table 2.

Funding: No funding is received for this work.

Conflict of Interest: The authors declare no conflict of interest.

References

[1] Xu, X.; Qi, Z.; Ma, J.; Zhang, H.; Shan, Y.; Qie, X.; "BTS: A Bi-lingual Benchmark for Text Segmentation in the Wild". Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 2022.

[2] Zhu, Y.; Yao, C.; Bai, X.; "Scene text detection and recognition: recent advances and future trends". Front. Comput. Sci., 10(1): 19–36, 2016.

[3] Chen, J.; Li, B.; Xue, X.; "Scene Text Telescope: Text-Focused Scene Image Super-Resolution". Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 2021.

[4] Fu, K.; Sun, L.; Kang, X.; Ren, F.; "Text Detection for Natural Scene based on MobileNet V2 and U-Net". Proceedings of the 2019 IEEE International Conference on Mechatronics and Automation (ICMA), Tianjin, China, 2019.

[5] Jaderberg, M.; Simonyan, K.; Vedaldi, A.; Zisserman, A.; "Reading Text in the Wild with Convolutional Neural Networks". Int. J. Comput. Vis., 116(1): 1–20, 2016.

[6] Long, S.; He, X.; Yao, C.; "Scene Text Detection and Recognition: The Deep Learning Era". Int. J. Comput. Vis., 129(1): 161–184, 2021.

[7] Feng, W.; Yin, F.; Zhang, X.-Y.; Liu, C.-L.; "Semantic-Aware Video Text Detection". In: Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 2021.

[8] Ye, Q.; Doermann, D.; "Text detection and recognition in imagery: A survey". IEEE Trans. Pattern Anal. Mach. Intell., 37(7): 1480–1500, 2014.

[9] Soni, R.; Kumar, B.; Chand, S.; "Text detection and localization in natural scene images based on text awareness score". Appl. Intell., 49(4): 1376–1405, 2019.

[10] Luo, C.; Lin, Q.; Liu, Y.; Jin, L.; Shen, C.; "Separating Content from Style Using Adversarial Learning for Recognizing Text in the Wild". Int. J. Comput. Vis., 129(4): 960–976, 2021.

[11] Raisi, Z.; Naiel, M.A.; Fieguth, P.; Wardell, S.; Zelek, J.; "Text Detection and Recognition in the Wild: A Review". arXiv, Jun. 30, 2020.

[12] Chen, X.; Jin, L.; Zhu, Y.; Luo, C.; Wang, T.; "Text Recognition in the Wild: A Survey". ACM Comput. Surv., 54(2): 1–35, 2022.

[13] Epshtein, B.; Ofek, E.; Wexler, Y.; "Detecting text in natural scenes with stroke width transform". In: Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 2010.

[14] Nistér, D.; Stewénius, H.; "Linear Time Maximally Stable Extremal Regions". In Computer Vision – ECCV 2008, Springer Berlin Heidelberg, vol. 5303, pp. 183–196, 2008.

[15] Bušta, M.; Neumann, L.; Matas, J.; "Deep TextSpotter: An End-to-End Trainable Scene Text Localization and Recognition Framework". In: Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017.

[16] Karatzas, D.; et al.; "ICDAR 2015 competition on Robust Reading". In: Proceedings of the 2015 13th International Conference on Document Analysis and Recognition (ICDAR), Tunis, Tunisia, 2015.

[17] Veit, A.; Matera, T.; Neumann, L.; Matas, J.; Belongie, S.; "COCO-Text: Dataset and Benchmark for Text Detection and Recognition in Natural Images". arXiv, Jun. 19, 2016.

[18] Kai, W.; Babenko, B.; Belongie, S.; "End-to-end scene text recognition". In: Proceedings of the 2011 International Conference on Computer Vision, Barcelona, 2011.

[19] Wolf, C.; Jolion, J.-M.; "Object count/area graphs for the evaluation of object detection and segmentation algorithms". Int. J. Doc. Anal. Recognit. IJDAR, 8(4): 280–296, 2006.

[20] Wang, W.; et al.; "Efficient and Accurate Arbitrary-Shaped Text Detection With Pixel Aggregation Network". In: Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), 2019.

[21] Liu, Y.; et al.; "Abcnet v2: Adaptive bezier-curve network for real-time end-to-end text spotting". IEEE Trans. Pattern Anal. Mach. Intell., 44(11): 8048–8064, 2021.

[22] Dinh, M.-T.; Choi, D.-J.; Lee, G.-S.; "Dense Text PVT: Pyramid Vision Transformer with Deep Multi-Scale Feature Refinement Network for Dense Text Detection". Sensors, 23(13): 5889, 2023.

[23] Neumann, L.; Matas, J.; "Real-time lexicon-free scene text localization and recognition". IEEE Trans. Pattern Anal. Mach. Intell., 38(9): 1872–1885, 2015.

[24] Mishra, A.; Alahari, K.; Jawahar, C. V.; "Top-down and bottom-up cues for scene text recognition". In: Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 2012.

[25] Shi, B.; Bai, X.; Belongie, S.; "Detecting Oriented Text in Natural Images by Linking Segments". In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017.

[26] Liu, Y.; Jin, L.; Zhang, S.; Luo, C.; Zhang, S.; "Curved scene text detection via transverse and longitudinal sequence connection". Pattern Recognit., 90: 337–345, 2019.

[27] Ghosh, M.; Mukherjee, H.; Obaidullah, S. M. Gao, X.-Z.; Roy, K.; "Scene text understanding: recapitulating the past decade". Artif. Intell. Rev., 56(12): 15301–15373, Dec. 2023.

[28] Neumann, L.; Matas, J.; "Real-time scene text localization and recognition". In: Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 2012.

[29] Xiangrong, C.; Yuille, A.L.; "Detecting and reading text in natural scenes". In: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2004), Washington, DC, USA, 2004.

[30] Liao, M.; Shi, B.; Bai, X.; Wang, X.; Li, W.H.; "Textboxes: A Fast Text Detector with a Single Deep Neural Network". In: Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA , 2017.

[31] Long, S.; Ruan, J.; Zhang, W.; He, X.; Wu, W.; Yao, C.; "TextSnake: A Flexible Representation for Detecting Text of Arbitrary Shapes". In: Proceedings of the European Conference on Computer Vision (ECCV), San Francisco, CA, USA, 2018.

[32] Graves, A.; Schmidhuber, J.; "Offline handwriting recognition with multidimensional recurrent neural networks". Adv. Neural Inf. Process. Syst., 21, 2008.

[33] Hochreiter, S.; Schmidhuber, J.; "Long short-term memory". Neural Comput., 9(8): 1735–1780, 1997.

[34] 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. Pattern Anal. Mach. Intell., 39(11): 2298–2304, 2016.

[35] Graves, A.; Fernández, S.; Gomez, F.; Schmidhuber, J.; "Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks". In: Proceedings of the 23rd International Conference on Machine Learning - ICML '06, San Francisco, CA, USA, 2006.

[36] Powers, D.M.W.; "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation." arXiv, Oct. 10, 2020.

[37] Sokolova, M.; Lapalme, G.; "A systematic analysis of performance measures for classification tasks". Inf. Process. Manag., 45(4): 427–437, 2009.

[38] Karatzas, D. et al.; "ICDAR 2013 Robust Reading Competition". In: Proceedings of the 2013 12th International Conference on Document Analysis and Recognition (ICDAR), Washington, DC, USA, 2013.

[39] Chen, H.; Tsai, S. S.; Schroth, G.; Chen, D. M.; Grzeszczuk, R.; Girod, B.; "Robust text detection in natural images with edge-enhanced Maximally Stable Extremal Regions". In: Proceedings of the 2011 18th IEEE International Conference on Image Processing, Brussels, Belgium, 2011.

[40] Yin, X.-C.; Pei, W.-Y.; Zhang, J.; Hao, H.-W.; “Multi-Orientation Scene Text Detection with Adaptive Clustering”. IEEE Trans. Pattern Anal. Mach. Intell., 37(9): 1930–1937, Sep. 2015.

[41] Baek, Y.; Lee, B.; Han, D.; Yun, S.; Lee, H.; "Character Region Awareness for Text Detection". In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019.

[42] Ch'ng, C. K.; Chan, C. S.; "Total-Text: A Comprehensive Dataset for Scene Text Detection and Recognition". In: Proceedings of the 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Kyoto, Japan, 2017.

[43] Xu Y., Wang Y., Zhou W., Wang Y., Yang Z., and Bai X.; "Textfield: Learning a deep direction field for irregular scene text detection". IEEE Trans. Image Process., 28(11): 5566–5579, 2019.

[44] Xing, L.; Tian, Z.; Huang, W.; Scott, M.; "Convolutional Character Networks". In: Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), 2019.

[45] Tian Z., Huang W., He T., He P., and Qiao Y.; "Detecting Text in Natural Image with Connectionist Text Proposal Network". In: Leibe B., Matas J., Sebe N., and Welling M. (eds) Computer Vision – ECCV 2016. Springer, Cham, pp. 56–72, 2016.

[46] Zhu, Y.; Chen, Liang, L.; Kuang, Z.; Jin, L.; Zhang, W.; "Fourier Contour Embedding for Arbitrary-Shaped Text Detection". In: Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 2021.

[47] Lyu, P.;, Liao, M.; Yao, C.; Wu, W.; Bai X.F.; "Mask TextSpotter: An End-to-End Trainable Neural Network for Spotting Text with Arbitrary Shapes". In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 67–83, 2018.

[48] Liu, Y.; Chen, H.; Shen, C.; He, T.; Jin, L.; Wang, L.; "ABCNet: Real-Time Scene Text Spotting With Adaptive Bezier-Curve Network". In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020.

[49] Liao, M.; Shi, B.; Bai, X.; "Textboxes++: A single-shot oriented scene text detector". IEEE Trans. Image Process., 27(8): 3676–3690, 2018.

[50] He, P.; Huang, W.; He, T.; Zhu, Q.; Qiao, Y.; Li, X.; "Single Shot Text Detector with Regional Attention". In: Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017.

[51] Ma, J.; et al.; "Arbitrary-oriented scene text detection via rotation proposals". IEEE Trans. Multimed., 20(11): 3111–3122, 2018.

[52] Sun, Y.; Zhang, C.; Huang, Z.; Liu, J.; Han, J.; Ding, E.; "TextNet: Irregular Text Reading from Images with an End-to-End Trainable Network". In: Jawahar C. V., Li H., Mori G., and Schindler K. (eds) Computer Vision – ACCV 2018. Springer, Cham, pp. 83–99, 2019.

[53] Zhou, X.; et al.; "EAST: An Efficient and Accurate Scene Text Detector". In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017.

[54] Cheng, Z.; Bai, F.; Xu, Y.; Zheng, G.; Pu, S.; Zhou, S.; "Focusing Attention: Towards Accurate Text Recognition in Natural Images". In: Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017.

[55] Li, H.; Wang, P.; Shen, C.; "Towards End-to-End Text Spotting with Convolutional Recurrent Neural Networks". In: Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017.

[56] Gupta, A.; Vedaldi, A.; Zisserman, A.; "Synthetic Data for Text Localisation in Natural Images". In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016.

[57] Sung, M.-C.; Jun, B.; Cho, H.; Kim, D.; "Scene text detection with robust character candidate extraction method". In: Proceedings of the 2015 13th International Conference on Document Analysis and Recognition (ICDAR), Tunis, Tunisia, 2015.

[58] Diaz-Escobar, J.; Kober, V.; "Text detection in natural scenes with phase congruency approach". In: Tescher A. G. (ed.) Applications of Digital Image Processing XL, SPIE, p. 115, 2017.

[59] Diaz-Escobar, J.; Kober, V.; "Natural scene text detection and recognition with a three-stage local phase-based algorithm". In:

Tescher A. G. (ed.) Applications of Digital Image Processing XLI, SPIE, p. 7, 2018.

[60] Turki, H.; Ben Halima, M.; Alimi, A. M.; "Text detection in natural scene images using two masks filtering". In: Proceedings of the 2016 IEEE/ACS 13th International Conference of Computer Systems and Applications (AICCSA), Agadir, Morocco, 2016.

[61] Chidiac, N.-M.; Damien, P.; Yaacoub, C.; "A robust algorithm for text extraction from images". In: Proceedings of the 2016 39th International Conference on Telecommunications and Signal Processing (TSP), Vienna, Austria, 2016.

[62] Dai, K.; Lu, J. ; Ruan, S.; "A Novel Method for Text Detection in Arbitrary Scenes Based on Multi-scale Segmentation Networks". In: Proceedings of the 2019 Chinese Control And Decision Conference (CCDC), Nanchang, China, 2019.

[63] Turki, H.; Ben Halima, M.; Alimi, A. M.; "A Hybrid Method of Natural Scene Text Detection Using MSERs Masks in HSV Space Color". In: Proceedings of the Ninth International Conference on Machine Vision, SPIE, p. 1034111, 2017.

[64] Zhang, S.-X.; Zhu, X.; Chen, L.; Hou, J.-B.; Yin, X.-C.; "Arbitrary Shape Text Detection via Segmentation with Probability Maps". arXiv, 2022.

[65] Dong, L.; Chao, Z.; Wang, J.; "An Efficient Detection Method for Text of Arbitrary Orientations in Natural Images". In: Yang C., Virk G. S., and Yang H. (eds) Wearable Sensors and Robots. Springer Singapore, pp. 447–460, 2017.

[66] Phuoc Huynh; Cong, et al; "CRAFT: Complementary Recommendation by Adversarial Feature Transform.". In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops, Munich, Germany, 2018.