Artificial Intelligence Review (2023) 56:15301–15373
https://doi.org/10.1007/s10462-023-10530-3

Scene text understanding: recapitulating the past decade

Mridul Ghosh[1,2] · Himadri Mukherjee[3] · Sk Md Obaidullah[2] · Xiao-Zhi Gao[4] · Kaushik Roy[3]

Published online: 18 June 2023

Abstract
Computational perception has indeed been dramatically modified and reformed from handcrafted feature-based techniques to the advent of deep learning. Scene text identification and recognition have inexorably been touched by this bow effort of upheaval, ushering in the period of deep learning. It is an important aspect of machine vision. Society has seen significant improvements in thinking, approach, and effectiveness over time. The goal of this study is to summarize and analyze the important developments and notable advancements in scene text identification and recognition over the past decade. We have discussed the significant handcrafted feature-based techniques which had been regarded as flagship systems in the past. They were succeeded by deep learning-based techniques. We have discussed such approaches from their inception to the development of complex models which have taken scene text identification to the next stage.

Keywords Scene text identification · Scene text recognition · Handcrafted features · Deep learning

✉ Kaushik Roy
kaushik.mrg@gmail.com

Mridul Ghosh
mridulxyz@gmail.com

Himadri Mukherjee
himadrim027@gmail.com

Sk Md Obaidullah
sk.obaidullah@gmail.com

Xiao-Zhi Gao
xiao-zhi.gao@uef.fi

[1] Department of Computer science, Shyampur Siddheswari Mahavidyalaya, Howrah 7113112, West Bengal, India

[2] Department of Computer Science, & Engineering, Aliah University, Kolkata, West Bengal, India

[3] TISA Lab, Department of Computer Science, West Bengal State University, Barasat, India

[4] School of Computing, University of Eastern Finland, Kuopio, Finland

1 Introduction

Text is a sequence of letters that portrays information in a document. This is one of the oldest methods of storing information and has been used to pass on information across ages. It has served an essential part in contemporary existence as being part of humankind's greatest impactful innovations. Text's extensive but accurate contextual knowledge is critical in disparate avenues like image search, target localization, human communication, robot control, and corporate management, because of the extensive and specific knowledge it contains. As a result, automated text identification and recognition, which allows users to access and use textual data in pictures and videos, has emerged as an important area in a variety of vision-based applications.

The goal of text detection is to locate the text within an input image, and the location is frequently displayed by a bounding box which is often rectangular in form. The objective of text recognition is to turn text-filled picture areas into sequences of system-comprehensible characters/words. Text detection is generally the first stage in text recognition. The task of identifying scripts of detected text can generally be done before text recognition or using the OCR (optical character recognition) technique for multi-script images. The script is a visual portrayal of literature that may be expressed using words or letters. This indicates that script is a component of communication, and in multi-script images, there are texts written in multiple scripts that may be inherited by several languages. English and German, for instance, use the Latin script, while Hindi and Marathi use the Devanagari script. Along with frequently featuring disparate font styles, colors, dimensions, and alignments, decorative/artistic or custom-designed font styles, complicated background graphics, foreground-background texture homogeneity, etc., are used in the display board, advertisements on hoarding, banners, electronic signboard, etc., for human attention. Different softwares are used for text recognition, especially for printed text documents like FreeOCR,[1] SimpleOCR,[2] GOCR,[3] Easy Screen OCR,[4] Tesseract,[5] ABBYY FineReader,[6] and PyPDF2 (Tesseract Based),[7] etc., still, they are yet to give satisfactory results in case of scene text image. Different scene text images in multifarious scenarios are shown in Fig. 1.

There are different categories of problems in scene text localization, identification, and recognition which are as follows:

- Scene images can have highly complicated backgrounds. Banners, hoardings, walls, meadows, etc., contain unique/ genuine writing styles, making them prone to cause misunderstandings or mistakes in automated recognition.
- Component proximity, connected letters, broken letters with several linked elements, combined characters, calligraphic texts, quasi text, range of typefaces, non-uniform/ artistic strokes, etc. are involved in the natural scene image texts.
- Several hindrances including noise, blur, distortion, poor resolution, in-homogeneous lighting, and incomplete occlusion, can cause text detection and identification errors.

[1] http://www.freeocr.net/ as visited on 01.11.2022.
[2] https://www.simpleocr.com/ as visited on 01.11.2022.
[3] https://jocr.sourceforge.net/ as visited on 01.11.2022.
[4] https://easyscreenocr.com/ as visited on 01.11.2022.
[5] https://tesseract-ocr.github.io/tessdoc/Downloads.html as visited on 01.11.2022.
[6] https://pdf.abbyy.com/ as visited on 01.11.2022.
[7] https://pypi.org/project/PyPDF2/ as visited on 01.11.2022.

Fig. 1 Different scene text images: **a** names of some cracker boxes written creatively in different orientations; **b** name of the shop is in multi-script and some portion is occluded; **c** hoarding where the name of a municipality office is present in four disparate scripts; **a** funky style of writing in t-shirt where a tagline is written with three scripts and a word is there which is in multi-script at character level; **a** name of a company in the circular inscription; **f** image of a title of a book which suffers from the reflection of light issue; **g** Tollywood movie poster with an artistic title; **h** a Hollywood movie poster where the title is written in Devanagari with homogeneous foreground-background

- Apart from disparate colors, and font sizes, the major challenge occurs when the text in images is non-horizontal or orientated, curved, and in transparent backgrounds.

1.1 Application scenarios

This section discusses a few of the uses for text detection and recognition in many fields, spanning from narrow systems to adaptive platforms.

- Recently, image processing-based apps are popular on various portable electronic devices that are built up with intelligence to help travelers who are unfamiliar with regional writing. Such an individual may be able to access crucial details from the intricate image they filmed. Also, the text-to-voice converter can help to understand more about the retrieved information from the cluttered scene.
- To determine a residence's number, a residence number plate is helpful. The number can be inscribed in a variety of typefaces, widths, and backdrop patterns. Systems focused on text data processing might make it easier to find and localize the image's textbox and recognize the text.
- Business houses can find applications when the requirement of the vast quantity of information from photos and recordings is to be extracted and interpreted automatically.

- Numerous sign panels and cautions are posted on the roadway to control congestion. The writing on such notices and signage conveys certain important signals to drivers and pedestrians. Such alerts and text messages are recognized and utilized to inform the public on how to lower the likelihood of traffic fatalities.
- Often displays are employed to distribute certain valuable messages for advertising or commercial purposes. A display includes a variety of elements, including integrated text, background graphics, images, and details about the item or vendor, etc., to catch human attention. Owing to different aspects, these texts are typically difficult to locate automatically.
- For robotic navigation, to perceive the scene/environment, real-time automated scene text analysis is a pressing need.

  In Table 1 several industrial applications of scene text detection and recognition are discussed.

The following are the main contrasts between the present survey and previous ones (shown in Table 2) and the main contributions of this work. Figure 3 depicts the common approaches to the analysis of scene text processing in the existing literature.

The survey is organized as follows: in Sect. 2 the methodology of selection of papers for this survey is introduced; the discussion on datasets is described in Sect. 3; in Sect. 4 the state-of-the-art is presented; in Sect. 5 observation is discussed which is followed by future scope (Sect. 6) and the survey is concluded in Sect. 7. For a better understanding, the organization of this study is presented diagrammatically in Fig. 2.

2 Survey methodology

The preferred reviews and meta-analyses (PRISMA) approach was followed to prepare and summarize this strategic literature survey. Figure 4 depicts the entire PRISMA investigation.

2.1 Standards for identification

A comprehensive search is done on web resources/depositories like Google Scholar, IEEE-Xplore, ResearchGate, springer, DBLP, and MDPI, etc. to collect relevant publications in an organized manner. For text detection and script identification in scene images, numerous questionnaires were utilized to improve the search performance. The following are a few of the most common data-gathering searching phrases in text detection in scene images: “scene text detection”, “deep learning-based scene text detection”, “classifiers used in scene text detection”, “text extraction in natural scene images”, “deep learning-based text extraction in the natural scene” etc. The following phrases were used in script identification-related papers: “script identification in wild”, “script classification in scene images”, “handcrafted-based methods in script identification”, “script identification in scene images using deep learning”, and “script identification in natural text images”. The publications in which the titles don’t reflect the localization and identification of scene text-related images are considered in the superfluous category and were removed from the collection set.

Table 1 Application scenarios of scene text detection and recognition techniques in different industries

Industries	Task	Applications
Tourism	Text detection, recognition	Automatic text understanding from traffic navigation, signboard, billboard, hoarding of different places, etc
Logistics and transportation	Automatic number plate recognition	Automatic toll collection, traffic rule violation detection, etc
Financial sector	Recognition/ identification of signature, legal amount, etc	Automatic check processing
Healthcare	Automatic document processing	Automatic patient history generation, prescription, and other healthcare documents extraction and storing
Retail	Extraction and processing of fields from various forms and print on packages	Records from invoices, shipping documents, bills, and customer orders extraction
Manufacturing	Extraction and processing of texts from various printing on packages	Automatic track of food and beverage processing from the raw material stage to packaged product

Table 2 Key differences of the present work with existing ones

Chen et al. (2021)	1. There is inadequate information regarding the criteria for the selection of the discussed works
	2. Phase-wise development of text detection, script identification, and text recognition is not described
	3. Results are not analyzed in depth
	4. They emphasized mainly recognition
Long et al. (2021)	1. Conventional machine learning-based works are less focused
	2. They emphasized mainly the deep Learning based approaches
Khan et al. (2021)	1. The pre-processing techniques are not highlighted
	2. They emphasized mainly the deep Learning based approaches
Lin et al. (2020)	1. Methods related to script separation in the multi-script environments to act as the precursor for text recognition didn't pay attention
	2. Only limited works are considered
Proposed	1. **This study chronologically focuses on text localization, script identification, and text recognition techniques in two phases i.e., from 2000–2012 and 2013–2021.**
	2. **A comprehensive categorization is presented with a broad analysis of available handcrafted feature-based techniques**
	3. **This study also emphasizes text analysis in natural images utilizing deep learning algorithms from initial to advanced phases**
	4. **It also highlights the key aspects of benchmark datasets along with in-depth exploration**
	5. **The results of the reported works are presented and discussed in detail for better understanding of the readers**

Bold is made to show the contribution of this paper

2.2 Standards of selection

The papers were selected based on the screening of keywords and abstract. The consideration criteria were set as:

1. Handcrafted-based techniques for text extraction,
2. Script separation using conventional machine learning-based techniques,
3. Text area prediction using deep-learning frameworks,
4. Deep learning-based techniques in text extraction and script identification,
5. Transfer learning-based approaches in scene text analysis.

After inspection based on these measures, the inadmissible publications were eliminated.

2.3 Standards of admissibility

The dataset details were noted based on the different methodologies employed in text extraction, script identification, and/or text recognition/analysis. The publications were grouped based on the datasets used in competitions, whether handcrafted-based,

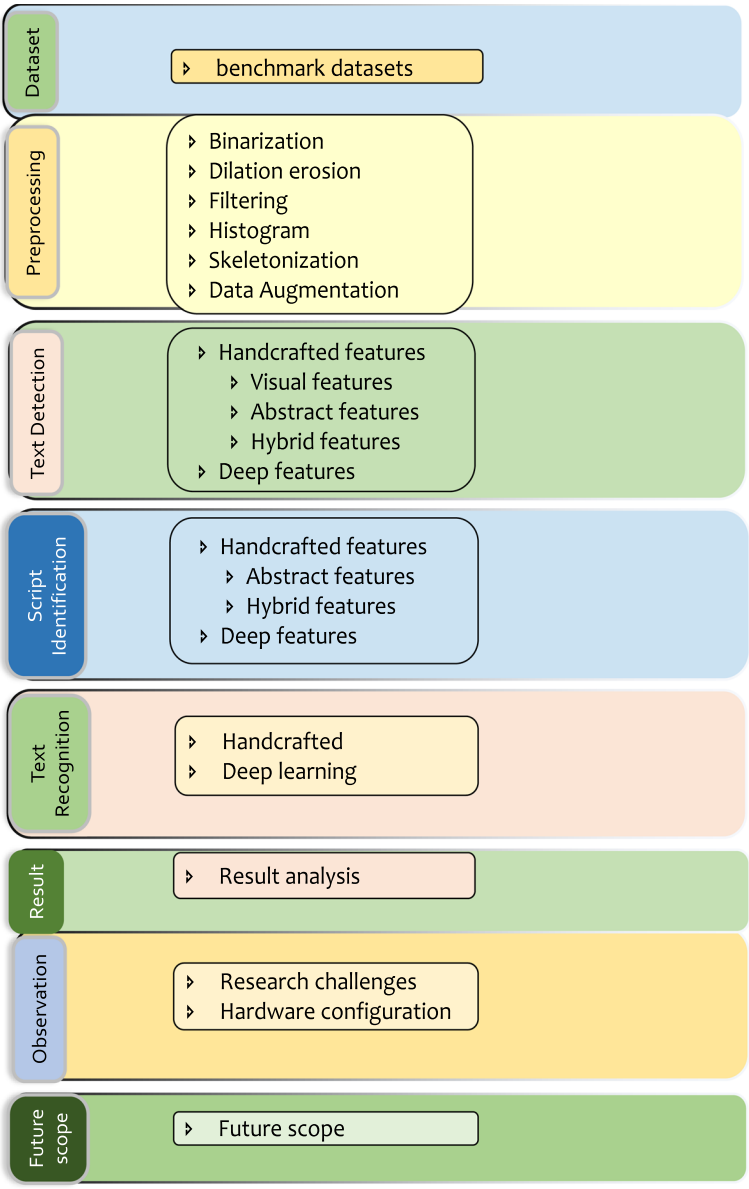

Fig. 2 The organization of this research

deep learning-based or hybrid techniques are used. The publications were ordered in sequence by considering the year-wise track of the methods to focus on the involvement of the progress. The publications were studied rigorously and the works were

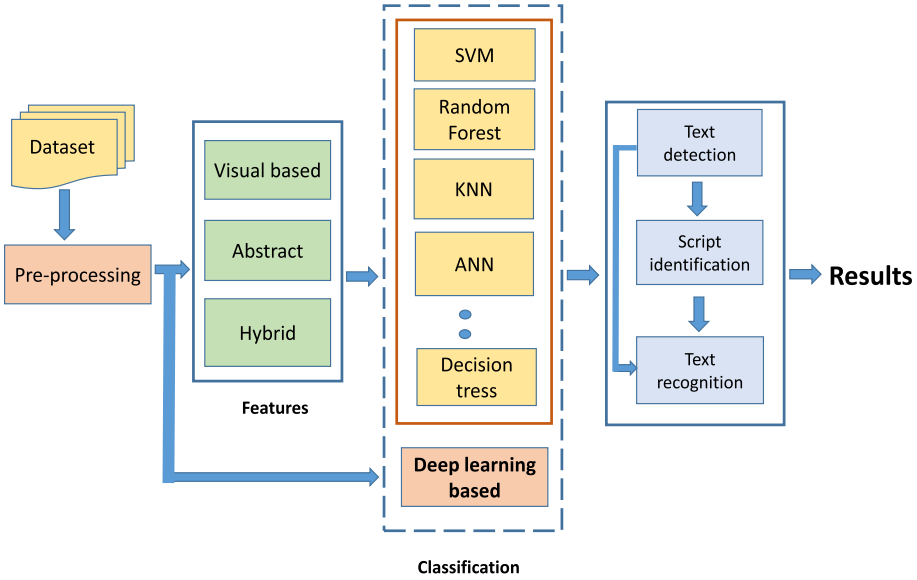

Fig. 3 The block diagram of the common approaches used in scene text analysis

summarized and analyzed. Finally, a report is prepared based on the observations achieved from the related works.

3 Benchmark datasets

Scene image-text recognition is not trivial for computers. To facilitate research in this avenue, different scene text image datasets have been proposed. The datasets have been composed of both low and high-resolution images as well as texts of multifarious orientations. Some datasets also consisted of images that suffered from faulty camera alignment, perspective deformation, low contrast, brightness, etc. In Table 3 the details of a few scene text image datasets are presented. This table contains the dataset along with their publishing year, specifications, and the text positions in the images.

3.1 ICDAR

There are different versions of scene text image datasets published in the ICDAR workshops as competition datasets. The ICDAR 2003 dataset is a robust reading competition that has issued its first baseline for scene text identification and recognition. There are 258 training images and 251 testing images in which the text is written in Roman. In the ICDAR2005 dataset, there are 258 training and 251 testing for character recognition purposes. The 2011 dataset was made of born-digital images i.e., collected from internet sources like webpages, emails, etc. In this set, there are 420 images where 3583 words from that were considered as training set and from 102 images, 918 words as testing. ICDAR2013 called focused scene text dataset in the RRC category contains 848-word

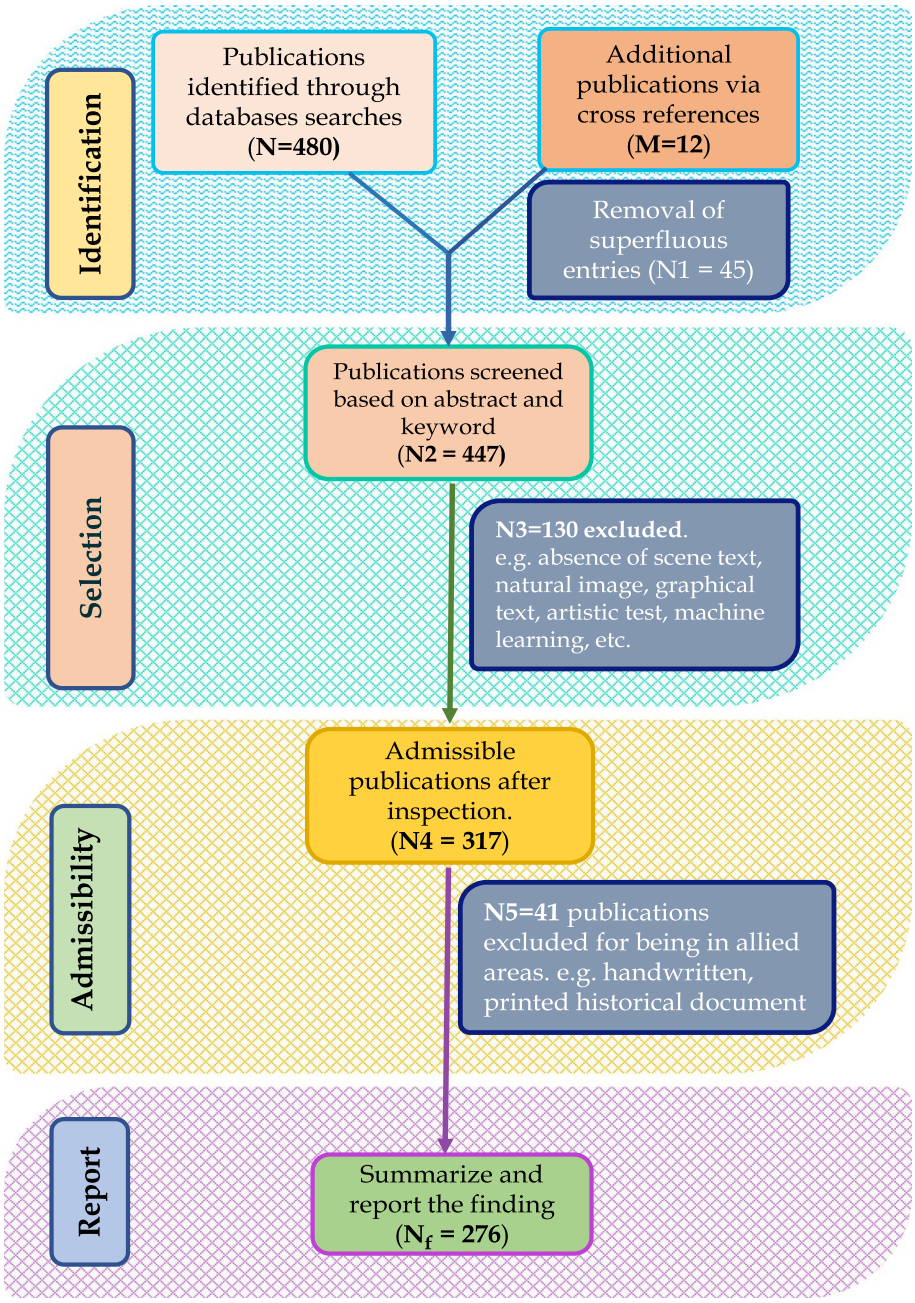

Fig. 4 The process of identification, selection, and analysis of publications

images for training and 1095 for testing. The scene images were captured by the commercial camera and the text is written in Roman. ICDAR 2015 RRC 1000 images for a train set and 500 images for a test set. In addition to the tasks on born-digital images, focused

Table 3 Datasets used in scene text analysis

Dataset name	Publishing year	Specifications	Text position
ICDAR 2003 (Lucas et al. 2005)	2003	Text locating: 258 images training, 251 testing; word recognition: 1157 words-training 1111 words-testing; character recognition: 6185 characters-training, 5430-testing	horizontal
ICDAR 2005 (Lucas 2005)	2005	258 training, 251 testing	Horizontal
Chars74K (Liu et al. 2019)	2009	7705 scene images	Horizontal
SVT (Mishra et al. 2012b)	2010	101 training, 249 testing	Horizontal
ICDAR 2011 (Shahab et al. 2011)	2011	Among 484 images 3583 words as training, 102 images 918 words as testing	Horizontal
IIIT 5K-word (Mishra et al. 2012a)	2012	1120 images, 5000 words	Horizontal, oriented
MSRA-TD500 (Yao et al. 2012)	2012	300 training, 200 testing	Horizontal
SVT-Perspective (Phan et al. 2013)	2013	238 images, 639 cropped words	horizontal, oriented, curved
ICDAR 2013 (Karatzas et al. 2013)	2013	229 training, 233 testing	Horizontal
CUTE80 (Risnumawan et al. 2014)	2014	288 cropped images	Oriented
ICDAR 2015 (Karatzas et al. 2015)	2015	1000 training, 500 testing	Arbitrary oriented
SIW-13 (Shi et al. 2016a)	2015	9791 training, 6500 testing	Horizontal
CVSI-15 (Sharma et al. 2015)	2015	6412 training, 3207 testing	Horizontal
Incidental scene text (Yao et al. 2015)	2015	1000 images, 4500 readable words	Horizontal, oriented, tilt, blur
SynthText (Gupta et al. 2016)	2016	85,8750 images	Horizontal
ICDAR 2017 MLT (Nayef et al. 2017)	2017	7200 training, 10,800 testing	Arbitrary oriented
COCO-Text (Veit et al. 2016)	2017	43,686 training, 20,000 testing	Arbitrary, Horizontal
ISI-UM (Dey et al. 2017)	2017	500 images	horizontal, oriented
MLe2e (Gomez et al. 2017)	2017	1178 training, 643 testing	Horizontal, oriented
RCTW-17 (Shi et al. 2017b)	2017	15,114 training, 1000 testing	Arbitrary, multi- oriented
Total-Text (Ch'ng and Chan 2017)	2017	1555 images, 9330 annotated words	Horizontal, multi- oriented, curved
CTW1500 (Yuliang et al. 2017)	2017	1500 words images	Arbitrary, horizontal
SCUT-CTW1500 (Yuliang et al. 2017)	2017	1000 training, 500 testing	Curved
ICDAR 2019 Arbitrary Text (Art) (Chng et al. 2019)	2019	5603 training, 4563 testing	Arbitrarily oriented

Table 3 (continued)

Dataset name	Publishing year	Specifications	Text position
DAST1500 (Tang et al. 2019)	2019	1038 Training, 500 testing	Horizontal, curved
ICDAR 2019 MLT (Nayef et al. 2019)	2019	10,000 training, 10,000 testing	Arbitrary oriented
ILST (Mathew et al. 2017)	2021	Around 3000 cropped word	Horizontal, oriented

scene images, and video text, a new challenge on incidental scene text has been considered in this set. In ICDAR 2017 MLT dataset consists of 68,613 training and 97,619 testing and 16,255 validation data of six scripts (Arabic, Latin, Chinese, Japanese, Korean, and Bangla). ICDAR 2019 MLT dataset contains 10,000 training and 10,000 test images of seven scripts. The multi-oriented scene texts are annotated using quadrangle boxes in this set. In ICDAR 2019 Arbitrary-Shaped Text (ArT19) there are 10,166 images, with 5603 for training and 4563 for testing. These were gathered with text form variety in view, and all text styles, including horizontal, multi-oriented, and bending, have a large number of examples.

3.2 Street view text (SVT)

This dataset comprises 350 images from Google Street View that were labeled as anchored boxes for word-level positioning. It contains tiny, low-resolution texts and not all of the text instances were marked.

3.3 SVT-perspective

Among 238 images, 639 cropped texts were extracted for testing in this set. There are potentially significant viewpoint abnormalities found in this dataset as the images were collected from Google Street View technology.

3.4 IIIT 5K-word

Due to the complicated surroundings, typeface, existence of distortion, and illuminated concerns, the IIIT 5K-Word dataset is a sizable and demanding dataset for benchmarking. There are 5000 photos captured and born-digital photographs in this standard dataset. Among them, 2000 were utilized for training and 3000 for testing.

3.5 Incidental scene text

This dataset refers to the capture of images in a variety of situations employing portable cameras in which the recording is hard to manage. Here, 4500 words are extracted from 1000 images.

3.6 SynthText

It comprises 858,750 synthetic images where the text components are of disparate fonts, background textures, colors, sizes, and orientations. The text segments were annotated in the line, word, and even at the character level.

3.7 MSRA-TD500

This dataset covers 500 scene images of Roman and Chinese text with 300 Training and 200 images for testing. The text line and words are annotated by polygon bounding boxes. COCO-Text: This benchmark dataset consists of 43,686 images for training and 20,000 images for testing that is used for text detection and recognition.

3.8 Chars74K

The Chars74K dataset is used to test character recognition techniques in real photos. This collection of characters includes English and Kannada letters. There are 7705 characters in this set, obtained from natural images.

3.9 ISI-UM

ISI-UM dataset was created in ISI Kolkata, India which contains 500 scene images for Bangla text detection.

3.10 IIIT ILST

It contains around 1000 word images, extracted from the scene images of each script Devanagari, Malayalam, and Telugu. The images were captured from diverse scenarios, such as neighborhood sectors of the economy, posters, route planning boards, road signboards, advertising, artwork, etc.

3.11 CVSI-15

This dataset contains 10,688 cropped word video frames having ten scripts of Bangla, Arabic Devanagari, Roman, Oriya, Gurumukhi, Tamil, Gujarati, Kannada, and Telugu.

3.12 MLe2e

In this dataset there are 1178 word images for training and 643 for testing of Chinese, Hangul, Latin, and Kannada scripts.

3.13 RCTW-17

This is a challenging dataset on identifying and recognizing Chinese text in photographs, comprising a variety of photos, spanning street views, billboards, restaurant menu cards, etc. There are 8000 training and 4000 test images in the dataset. Total-Text: The dataset comprises street signboard images having a broad range of horizontal, curved, and multi-oriented labels using polygonal bounding box coordinates at the word level. It consists of 1555 scene images with 9330 annotated words.

3.14 CTW1500

It comprises 1500 images, each with minimum single bend text. There are 3530 bend words in the 10,751 enclosing word containers. The images were collected from the web, and image libraries, and captured by phone camera which included lateral and multi-oriented lettering.

3.15 DAST1500

It comprises 1538 images where the text is arbitrarily oriented. There are 1038 training and 500 testing images.

3.16 Total-text

There are 1555 photos inside the Total-Text dataset, containing 11,459 truncated text specimen images. There are more than three possible directions for images in Total-Text, namely horizontal, multi-oriented, and curved.

3.17 SCUT-CTW1500

With 10,751 truncated word images, this dataset has 1500 images in the aggregate, 1000 for training, and 500 for testing. CTW-1500 annotations are polygons with 14 nodes. The majority of the words in the sample are Chinese and English.

3.18 CUTE80

It includes 80 bent text images with a complicated background, viewpoint deformation impact, and low-quality implications. The images were captured by the camera and collected via the web.

3.19 SIW-13

In this dataset there are 16,291 text images collected from Google street view technology having 13 scripts of Greek, Mongolian, Cambodian, Tibetan, Thai Arabic, Chinese, Kannada, English, Korean, Hebrew, Japanese, and Russian.

Figure 5 shows sample scene text images from several datasets.

4 State-of-the-art

The developments in scene text processing in the past decades are broadly classified into handcrafted and deep learning-based methods. For ease of understanding, they are further divided into 2 chronological categories: phase 1 and phase 2. For handcrafted feature-based

Fig. 5 Sample images from some of the public datasets. One sample image from **a** ICDAR 2003 (Lucas et al. 2005), **b** ICDAR 2005 (Lucas 2005), **c** ICDAR 2011 (Shahab et al. 2011), **d** ICDAR 2013 (Karatzas et al. 2013), **e** ICDAR 2015 (Karatzas et al. 2015), **f** ICDAR 2017 MLT (Nayef et al. 2017), **g** ICDAR 2019 MLT (Chng et al. 2019), **h** COCO-Text (Veit et al. 2016), **i** SVT (Mishra et al. 2012b), **j** IIIT 5K-word (Mishra et al. 2012a), **k** MSRA-TD500 (Yao et al. 2012), **l** Chars74K (Liu et al. 2019), **m** ILST (Mathew et al. 2017), **n** CVSI-15 (Sharma et al. 2015), **o** RCTW-17 (Shi et al. 2017b), **p** CTW1500 (Yuliang et al. 2017), **q** CUTE80 (Risnumawan et al. 2014), **r** SIW-13 (Shi et al. 2016a), **s** SynthText (Gupta et al. 2016), and **t** Total-text (Ch'ng and Chan 2017) datasets is shown

methods, phases 1 and 2 are considered from 2000–2012 and 2013–2021, while for deep learning-based methods phases 1 and 2 are studied from 2012–2016 and 2017–2021, respectively. These techniques are highly dependent on the different pre-processing methods, a few of which are discussed in the following section.

4.1 Pre-processing techniques

The text localization and recognition system's pre-processing phase rests on the capacity to resolve several issues like background complexity, color uniformity in foreground-background, capturing errors like the camera-sensor heating issue, blurriness, noisiness, etc. The pre-processing approaches can therefore improve the quality of natural images, providing good assistance for the text segmentation to the OCR engine in terms of the performance of the system. Many techniques were proposed for this purpose, few of them are presented as follows.

Binarization On colored/gray text images comprising text and/or visuals, processing procedures are required. Because analyzing colored images is algorithmically intensive.

Fig. 6 Progress in pre-processing of scene text image. The source image **a** represents a Hollywood movie poster, and **b** is taken from a bottle cover. The output of Otsu's binarization technique (Otsu 1979) are shown in (**c**, **d**) and the binarized images generated by an auto-encoder-based approach (Calvo-Zaragoza and Gallego 2019) is presented in (**e**, **f**) for the images shown in (**a**, **b**), respectively

Several letter identification algorithms utilize grey or monochrome images. To isolate an image's content from its background, an image binarization/thresholding operation is used generally. To reduce the number of channels and complexity of the system, Otsu's global binarization technique (Otsu 1979) was followed (Dhar et al. 2020; Ghoshal and Banerjee 2020; Ghosh et al. 2018; Sengupta and Mollah 2021) in scene image processing. There are other techniques in the literature (Howe 2011; Lu et al. 2010; Pratikakis et al. 2013; Wolf and Doermann 2002) for this process. Apart from the conventional approach, deep learning-based methods (Afzal et al. 2015; Pastor-Pellicer et al. 2015) were also considered. In Peng et al. (2017), Calvo-Zaragoza and Gallego (2019) authors proposed an auto-encoder-based binarization technique for document images.

In Fig. 6 the visual comparison of the Otsu's (1979) and auto-encoder-based (Calvo-Zaragoza and Gallego 2019) binarization techniques are shown. Figure 6b, c (by Otsu) and e, f (auto-encoder based) are the binarized form of the original scene images (a) and (b). It is obvious from the figure that the auto-encoder-based technique gives us good binarization performance.

Filtering The smoothing/filtering operation sometimes helps to improve the quality of the image. For this purpose, different filters were used in the literature. Authors used Sobel (Phan et al. 2012; Su et al. et al. 2019; Shinde and Patil 2021; Turki et al. 2016), Canny (Epshtein et al. 2010; Phan et al. 2011; Sravani et al. 2021; Yao et al. 2012), and Edge Preserving Smoothing Filter (EPSF) (Huang et al. 2013a; Soni et al. 2019), etc., to find out the contours of the text objects in the image. Shivakumara et al. (2012) applied the Sobel operation to suppress low-contrast pixels to enhance text detection performance. Yao et al. (2012) used a canny operator for edge detection in multi-oriented text detection in natural images. For text localization, Pan et al. (2010a) presented a

region filtering scheme to improve the sharpening of the edges of text sections. Feng et al. (2016) used an edge filtering-based approach to filtering out non-text pixels.

Histogram based From the histogram analysis it was seen that the textual rectangles were laterally aligned (Nagaoka et al. 2021; Ren and Ramanan 2013). The edge information (Hu et al. 2021) and visual features (Yi and Tian 2013) of alphabets/letters/words were extracted from the histogram. Authors in Simanjuntak and Nugroho (2021) improved the contrast of the images by using the histogram equalization technique.

Erosion-dilation Authors (Del Gobbo and Herrera 2020; Jang et al. 2002; Bhattacharyya et al. 2020; Sravani et al. 2021; Zharikov et al. 2020) used erosion and dilation methods to eliminate the text-like components which often mislead the system as text in the images. According to Dhar et al. (2020) the text contains higher or lower pixel values than the background. After dilation and erosion for both cases and performing a complementary operation, the text areas are found.

Skeletonization By deleting the majority of the raw foreground pixels, the foreground areas of a single-tone image can be reduced to structural remnants despite mostly preserving their length or connectedness. Authors (Azadboni et al. 2014; Agrawal and Varma 2012) used the skeletonization process for scene text detection and extraction.

Apart from conventional pre-possessing techniques for improving the quality of the images, data augmentation techniques were also adopted in the pre-possessing phase to synthetic data generation for increasing the volume of datasets and to improve the performance of the system (Atienza 2021a; Ghosh et al. 2021b, 2022).

Data augmentation:

Noise The images may be contaminated with noise at the moment of picture shooting due to poor illumination, high temperatures, unanticipated changes, etc. In augmentation, images are being contaminated with different types of noises like salt & pepper, Gaussian, Poisson, etc.

Rotation In various angles of rotation ϕ the scene images can be turned to make synthesized images.

Shearing Implementing shear causes the image's appearance to be twisted. It can be performed both horizontally and vertically.

Wavelet transform To work with images having disparate resolutions, to get the advantages of concurrent localization of frequency and time, fast processing of images, and images using less memory, wavelet transform can be used. The wavelet decomposition can be performed in different levels of approximate and detail coefficients.

Blur Often real images suffers from blurriness issue. In data augmentation, the blurry effect can be imposed by allowing the low frequency to join while stopping the high frequency using high-cut filters.

In Fig. 7 output images after different data augmentation techniques are presented.

4.2 Handcrafted feature-based systems

In these systems, the features are extracted from the images using different handcrafted machine learning-based algorithms, and the class categorization is performed by using various classifiers.

Fig. 7 Image augmentation: **a** Original image: Tollywood movie poster (Bangla), **b** 45°, **c** 90°, **d** 135° rotated, **e** sheared image having shearing value 0.45, **f** Gaussian noisy image of 20% density, **g** single-level Haar wavelet decomposed image, and **h** blurry

4.2.1 Phase 1 (2000–2012)

In this phase, researchers paid attention to the necessity to extract the texts from natural images more to ease in many applications like traffic navigation, content-based image retrieval, information processing, etc. Different methods like visual-based where the features extracted are by considering the topological structures of the characters; abstract-based where the texture features are considered; and hybridized where the amalgamation of visual and abstract-based or amalgamation of different individual abstract methods are considered. Different techniques like MSER (Maximally stable Extremal region), SWT (Stroke width transform), CRF (Conditional random field) based models, etc., were developed in this phase.

4.2.1.1 Text detection *Visual features:* A plethora of methods (Chen et al. 2011; Shivakumara et al. 2010; Shi et al. 2012; Yao et al. 2012; Yi and Tian 2011) were proposed to extract the text elements utilizing manually set criteria or dynamically learned algorithms. In the visual characteristic-based approach, letter geometry such as edges, lines, shapes, and corner points is considered for both text and no-text areas using color non-uniformity, contrast change, etc.

In 2000, Garcia and Apostolidis (2000) used the Deriche edge detector and clustering techniques to detect the text in natural images. Neumann and Matas (2010) highlighted a technique that addressed the shortcomings of state-of-the-art methods. They observed the extraction of the text areas using complex procedures and identification of the coordinate points for the problem of arbitrary text orientation. They sampled and grouped the corner points of the probable boxes and using the position-sensitive segmentation map the localized area text regions are extracted. Shivakumara et al. (2010) suggested a technique for a multi-oriented text identification system. A skeletonization-based approach was used to

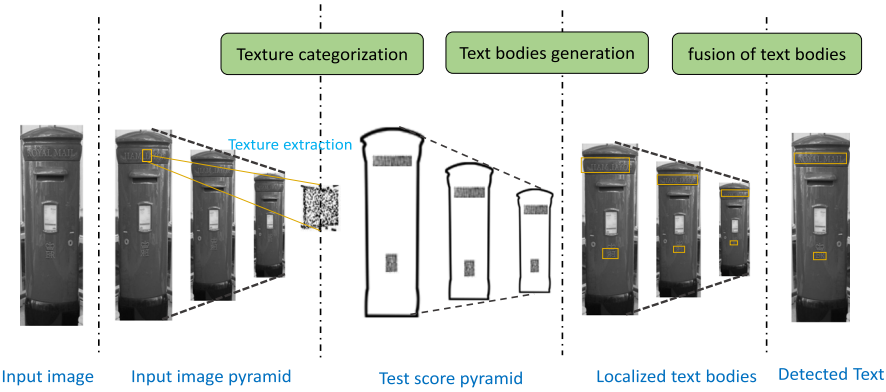

Fig. 8 The key process employed by the text detection system (Kim et al. 2003) to detect the presence of text using SVM and adaptive mean shift technique

split potential areas into different elements after grouping in the Fourier–Laplace space. Such elements do not usually correlate to strokes or letters, yet rather to text sections. Chen et al. (2011) also proposed MSER-based methods to extract text in scene images. In 2011, authors in Yi and Tian (2011) detected the text regions by considering the uniformity in color and local gradient-based features. They found out the connected components in characters from the features and formed character grouping to generate text strings. They experimented with ICDAR 2003 and their developed oriented scene text dataset. Chowdhury et al. (2011) proposed a text detection method using Euclidean distance measure, comparing the text and non-text portions. With the help of stroke thickness and Hough line transform, the header line characteristics of Bangla and Devanagari scripts are identified. Yao et al. (2012) presented a method for detecting words in natural pictures with variable orientations. This method includes a two-level categorization strategy as well as two sets of rotation and rotation-invariant attributes that were created specifically for acquiring the fundamental properties of characters in real scenarios. Gonzalez et al. (2012) used MSER and the adaptive thresholding method to segregate the non-text component from the image. Yin et al. (2012) proposed an automatic text localization technique using the MSER method. Disjoint set analysis and trained Adaboost classifiers were used to identify the text regions. Shi et al. (2012) combined pixel-level and context-level characteristics into a graph by treating every pixel as a vertex in the network to detect text in scene images.

Abstract features: In this approach different texture-based features are considered in scene text analysis. To differentiate between text and non-text portions or to identify the text areas in images, authors in Chen and Yuille (2004), Kim et al. (2003), Gllavata et al. (2004) considered texts as a particular sort of texture and used their textural features, like regional intensities, filter outputs, wavelet coefficients, etc. Because all regions and scales must be inspected, these approaches are generally algorithmically intensive. Furthermore, these approaches primarily deal with lateral texts and are responsive to fluctuations in size and orientation. Kim et al. (2003) used the actual pixel values as a feature descriptor to train the SVM (support vector machine) classifier to categorize each pixel. An adaptive mean shift likelihood map was used to find text regions. Their approach achieves a miss rate of 2.4% in text detection with naive perspectives, but this method is not suitable for

complicated natural scenes. Figure 8 provides an overview of the process involved in their text detection.

A refined multi-scale scanning technique was presented by Lyu et al. (2005) to identify the text regions. Along with features like sharp edges and a significant contrast of texts, they also considered the regional dynamic Boolean approach for segmenting the text regions. Epshtein et al. (2010) discussed a robust feature extraction technique for text detection, based on SWT by combining local and non-local scales. They showed that their method is fifteen times faster than state-of-the-art techniques. The text region extraction from natural images was discussed in Angadi and Kodabagi (2010) where DCT-based (discrete cosine transform) high-pass filters were deployed for background removal and texture-based methods were used for segmentation and detection of text areas.

A technique to extract specific phrases in natural scenarios as suggested by Wang and Belongie (2010). They used a rolling pane to identify single letters. Then, based on the architectural links between letters, potential permutations were rated. In the outcome findings, the most comparable permutations were chosen using the supplied collection. Despite typical text detection algorithms, this method could only recognize words from the provided collection and cannot handle words from outside the collection. Unfortunately, for each image, a sequence including all viable terms is rarely accessible. In comparison to existing text identification algorithms, doing so limits the technique's application. Kasar and Ramakrishnan (2011) considered geometric, boundary, gradient, and stroke-based features for text localization from the multi-oriented and multi-scripted text in natural images. To extract the genuine bounding boxes (Wang et al. 2012) considered text line identification by creating a response map using a sliding window over the image which is followed by NMS (non-max suppression) method. They integrated the localities of the bounding boxes for words as well as for characters using beam search to segment and identify the words in a line.

Hybrid features: These are an amalgamation of different abstract-based or a mixture of abstract-based and visual-based techniques that combine the benefits of individual methodologies. In 2000, Li et al. (2000) proposed a hybrid technique using wavelet and neural network-based approaches to detect text in digital video frames. Using a collection of textural characteristics [e.g. different descriptor modules of HOG (Histogram of Gradient)] (Dalal and Triggs 2005) calculated with a set of preset patterns to determine the likelihood layouts. To distinguish text parts from non-text regions, a Conditional Random Field (CRF) model (Pan et al. 2009) was also used, which combines unary element characteristics with bipolar spatial connections. By integrating the texture-based non-text region information with the connected components, a hybrid technique was proposed by Pan et al. (2010b) to identify text strings in scene images. On experimenting with the ICDAR2005 dataset (Lucas 2005), they found an improved f-score of 0.62. A hybridization of texture-based features extracted from wavelet histogram and HOG techniques was proposed by Darab and Rahmati (2012) to localize text. The summary of techniques and text detection results are depicted in Table 4.

4.2.1.2 Script identification The goal of script identification is to detect the script of a particular text. In multidisciplinary platforms, it is becoming highly crucial since text recognition can choose the proper language paradigm by identifying the script and language. Script identification may be viewed as an image categorization issue in which exclusionary forms, like mid-level characteristics, are often developed.

Table 4 The methods used and results reported for text detection are based on handcrafted features (Phase 1)

Author	Dataset	Technique	Performance
Garcia and Apostolidis (2000)	200 images	Deriche edge detector	A-93%
Li et al. (2000)	150 frames	Wavelet, neural network	A-92.40%
Lyu et al. (2005)	Video images	Edge detection, thresholding	A-90.80
Epshtein et al. (2010)	ICDAR 2003, ICDAR 2005	SWT	P-0.73, R-0.60, F-0.66, T-0.94 s
Neumann and Matas (2010)	ICDAR 2003	Hypotheses verification system, MSER	P-0.59, R-0.55, F- 0.57
Angadi and Kodabagi (2010)	100 images	DCT-based	A-96.60%
Kasar and Ramakrishnan (2011)	ICDAR 2003	Unsupervised clustering, geometric, boundary, stroke and gradient based features	P-0.8, F-0.86
Pan et al. (2010b)	ICDAR 2005	Text portion detector, CRF	P-67.4, R-69.7, F-68.5
Chowdhury et al. (2011)	120 images	Probabilistic Hough line transform	P- 0.72, R-0.74
Yin et al. (2012)	ICDAR 2011	MSER, topology based letters area cluster, classification using AdaBoost	P-62.22, R- 81.53, F-70.58
Darab and Rahmati (2012)	800 images	Edge, ColorHOG, Wavelet coefficient histogram	P-29.40, R-86.50, F-43.90; P-80.80, R-83.50, F-86.50
Yao et al. (2012)	MSRA TD-500,ICDAR2005	SWT	P-0.69, R-0.66, F-0.67; P-0.77, R-0.73,F-0.74
Neumann and Matas (2012)	ICDAR 2011	Extremal region detector	P-68.7, R-64.7, F-73.1
Yi and Tian (2011)	ICDAR2003, ICDAR 2011	GMM, expectation-maximization EM, stroke segregation, Gabor features	P-0.81, R-0.72, F-0.71
Chowdhury et al. (2011)	ICDAR2003	Connected component analysis by Canny detector, stroke thickness	P-0.72, R- 0.74
Gonzalez et al. (2012)	ICDAR203	MSER	P-0.81, R-0.57, F-0.67

Results on multiple datasets in the same work are presented in the same order as dataset names are presented. The accuracy, precision, recall, and f-measure/F-score are represented here by A, P, R, and F, respectively

Abstract features: To differentiate between Latin and Ideographic script in images with complicated surroundings, authors (Gllavata and Freisleben 2005) proposed an approach that makes decisions based on a collection of characteristics retrieved straight from the source image. Phan et al. (2011) considered video script identification based on the text segments. They applied a canny edge detector to identify the text segments. They extracted higher and lower extreme endpoints from the text lines to analyze the behavior of the top and bottom lines and the retrieved endpoints which are linked. They applied PCA (Principal Component Analysis) to identify the orientation of every ten-pixel component of the lines. A method of extracting spatial gradient features for video script identification was proposed by Zhao et al. (2012).

Hybrid features: De Campos et al. (2009) used a hybrid method by considering the shape and texture-based approach to recognize the characters in natural images. Shape Contexts(SC), Geometric Blur (GB), Scale Invariant Feature Transform (SIFT) extracted topological features and Maximum Response of filters (MR8), Patch descriptor (PCH), and Spin were exploited as texture-based features.

4.2.1.3 Text recognition Text recognition can be used to convert a shortened text occurrence into the desired string pattern. It's a crucial part of an end-to-end mechanism that delivers reliable recognition outcomes. Hand-crafted attributes like the histogram of directed slope descriptors, linked elements, stroke width morph, etc., are used in conventional text recognition algorithms.

A scaled attribute acquisition technique was discussed by Coates et al. (2011). They used bigger pools of attributes to maintain better efficiency compared to certain competing methods, which was related to what was shown across various fields like computer intelligence and computer vision. Neumann and Matas (2012) proposed an end-to-end model to localize and recognize scene text in real time. They presented the character recognition issue as an optimal progressive decision among a collection of extremal zones to deal with the issues related to non-uniform background, noise, blur, low brightness contrast, etc. In 2012, Mishra et al. (2012b) generated bottom-up signals from each component occurrence in the image. They developed a conditional random field method to holistically describe the intensity of the detection. Here lexicon-based priors were considered in top-down signals. In their other work (Mishra et al. 2012a) of text recognition of natural scene images, they used a higher-level probabilistic language model. They developed a large dataset of around five thousand images and named it the IIIT 5K-word dataset. In Table 5 the script identification and text recognition results are tabulated.

4.2.1.4 Classifiers Gllavata and Freisleben (2005) followed k-NN (k-nearest neighbor) classification scheme on Euclidean, Bhattacharyya, and Manhattan distance metrics. Phan et al. (2011) also used the K-NN classifier for script categorization. Zhao et al. (2012) used Euclidean distance for classification. De Campos et al. (2009) MKL (multiple kernel learning), K-NN, and SVM classifiers were used to separate the Roman and Kannada text images. Neumann and Matas (2012) regarded AdaBoost classifier on decision trees, SVM, and RBF (Radial Basis Function) in their real-time text recognition work. Coates et al. (2011) used L2-SVM for digit classification in natural images. Mishra et al. considered SVM classifier in their work (Mishra et al. 2012a, b) for text recognition. They Mishra et al. (2012b) are also considered SVM classifiers with an RBF kernel.

Table 5 Script identification and recognition using handcrafted-based methods in phase 1

Author	Dataset	Technique	Performance
Gllavata and Freisleben (2005)	Developed	Low level feature extraction, K-NN classifier	A-85.30, 89.00
Coates et al. (2011)	ICDAR 2003	K-means clustering variation	A-85.5
Phan et al. (2011)	English-200, Chinese-150, Tamil-150	Canny edge detector, PCA, K-NN	A-67.00, 60.00, 8.00 (50% data in testing); A-74.4, 60.7, 11.8 (90% data in testing)
Mishra et al. (2012b)	SVT, ICDAR 2003	Conditional random field	A-73.26, 81.78
Mishra et al. (2012a)	IIIT 5K-word	Graph based model	A-64.10
Neumann and Matas (2012)	SVT, ICDAR 2011	Shape based	R-64.7, P-73.1, F-68.7
Zhao et al. (2012)	Chinese-100, Arabic-100, English-260, Tamil-100, Japanese-100, Korean-100	Gradient Histogram	A-82.10
De Campos et al. (2009)	English-62 classes, Kannada-600 classes	Hybrid feature	A-55.26

4.2.2 Phase 2 (2013–2021)

In this phase, researchers dealt with the challenges inherent in analyzing the complex scene text images where the background and foreground color/texture differences are little. They experimented with images where along with horizontal text lines, oriented and curved text segments are present.

4.2.2.1 Text detection Text detection has been a long-standing challenge in machine learning and computer vision. We categorize the features for text detection in this phase into three ways: visual, abstract, and hybrid features.

Visual features: Gomez and Karatzas (2013) used synthetic fonts and a hypothesis-verification framework for parallel-processing multiple text lines and using the MSER technique to identify the text and non-text portions. Raghunandan et al. (2018) discussed text detection in scene image/video script by iterative nearest neighbor symmetry and mutual nearest neighbor pair methods. They used the SVM method to calculate the classification score. Authors in Xie and Tu (2015) proposed a holistically nested edge detection (HNED) method for object detection in images. In Ghosh et al. (2020) a semi-automated character segmentation was discussed. In this work, authors used connected component analysis for character segregation and a manual segmentation was done afterward for incorrect segmented characters.

Abstract features: In Yin et al. (2013) authors discussed a technique that is centered on the idea of pruning. They mined the characters by the MSER technique and clustering algorithm where path weights and thresholds are determined using their self-training distance metric method. These characters are grouped to form text regions. Using probability theory the score for the text and non-text section is calculated and based on this final text regions were inferred. Risnumawan et al. (2014) presented a comprehensive approach using MMS (Mutual Magnitude Symmetry), MDS (Mutual Direction Symmetry), and GVS (Gradient Vector Symmetry) features to find textual unit choices in real-world photos independent of alignment, curvature, etc. This approach was premised on the reality that the textual sequences obtained by the Sobel and Canny edge mappings inside the source pictures, behave similarly.

Yao et al. (2014a) suggested the extraction of features and classification schemes by modifying the Random forest classifier for multi-oriented natural images and using dictionary search-based techniques for error correction. Authors in Kumuda and Basavaraj (2015) explained how to extract text regions using texture-based features. They considered statistical analysis to extract features by a first-order probability distribution and GLCM (Gray-Level Co-occurrence Matrix) features and discriminative analysis was done for non-text region identification. In Shivakumara et al. (2015) Shivakumara et al. extracted handcrafted features at the block level using gradient special and gradient structural features to identify six video scripts. They used the skeletonization method to reduce the pixel width and took different points relating to pixels for feature consideration. Dey et al. (2017) proposed a text detection method using ring radius transform of multi-script and multi-oriented natural scene images. They experimented with public datasets like ICDAR 2013, SVT, and MSRA along with their developed Bangla dataset named ISI-UM. Apart from Bangla, they experimented with various scripts like Japanese, Chinese, Tamil, Korean, and Arabic. Based on the adaptive stroke filter and component labeling techniques, Paul et al. (2019) suggested scene text localization.

Table 6 The methods used and results reported for text detection using handcrafted feature-based methods (Phase 2)

Author	Dataset	Technique	Performance
Gomez and Karatzas (2013)	KAIST	MSER Group hypothesis	P-0.66, R-0.78, F-0.71,Time(s)-0.41
Yin et al. (2013)	ICDAR 2011	Pruning method over MSER, distance metric learning algorithm	P-68.26, R-86.29, F-76.22
Yao et al. (2014a)	ICDAR 2011, MSRA-TD500	SWT, clustering	P-0.822, R-0.657, F-0.730; P-0.64, R-0.62, F-0.61
Risnumawan et al. (2014)	ICDAR 2005, ICDAR 2011, MSRA-TD500	MMS, MDS, GVS, SIFT	P-0.76, R-0.63,F- 0.69, Time(s)-15.8; P-0.83, R-0.71, F-0.77, Time-13.9; P-0.70, R-0.68, F-0.69
Xie and Tu (2015)	BSDS500 NYUDv2	HNED	P-0.833, 0.786
Kumuda and Basavaraj (2015)	ICDAR 2011	Texture features	P-90.00
Shivakumara et al. (2015)	Arabic-200, Chinese-200, English-200, Japanese-200, Korean-200, Tamil-200	Gadient-based features	P-80.40
Dey et al. (2017)	ISI-UM, SVT, MSRA	RRT (Ring Radius Transform)	P-0.79, R-0.66, F- 0.72;P-0.68, R-0.55, F-0.61; P-0.85, R-0.52, F-0.65
Raghunandan et al. (2018)	ICDAR 2011, ICDAR 2013, ICDAR 2015, MSRA, YVT, SVT	INNS, MNNP	R-81.10, P-76.30, F-78.60; R-87.70, P-84.50, F-86.00; R-67.60, P-62.80, F-66.10; R-77.20, P-67.20, F-72.40; R-81.60, P-78.80, F-80.10, R-68.70, P-60.40, F-64.20
Paul et al. (2019)	600 images	Fuzzy distance transform stroke filter, MNNP	R-96.65, P-87.77, F-91.89

Hybrid features: A hybrid approach presented in Turki et al. (2017) constructed by three phases to segment out the text regions. In the first phase, Otsu's approach to restricting the text regions via modification and a strong edge projection screen of the complicated surroundings and MSER technique was used to recognize the potential text pixels. A heuristic filtering-enhanced SWT method was followed to trim the predicted letter contenders in the second phase. Then, to filter out non-text elements, the categorization was done by relying on the SVM classifier. In Rashmi and Nayak (2018) another hybrid approach was discussed using the standard filters like average, Prewitt, the edge features, and by exploiting GLCM, texture features were extracted to detect text regions in natural images. In Table 6 the brief description of literature works along with their results are tabulated.

4.2.2.2 Script identification Script identification is an important part of OCR, which has gotten a lot of interest in multi-script image processing.

Abstract features: Mid-level features (Fernando et al. 2014; Juneja et al. 2013) were considered in script identification which is based on the concept of feature extraction in Boureau et al. (2010). Singh et al. (2016) extracted mid-level features and demonstrated an end-to-end script identification workflow. They experimented with their proposed dataset named ILST (Mathew et al. 2017) along with the public dataset CVSI-15. Authors (Singh et al. 2016) proposed script identification of natural scene images by a mid-level illustration of SIFT feature descriptors on CVSI-15 and ILST datasets. They developed ILST (Indian Language Scene Text) dataset for this work. Verma et al. (2017) extracted texture-based features by using LBP (Local Binary Pattern), CS-LBP (Center Symmetric Local Binary Pattern), and DLEP (Directional local extrema pattern) to identify the multi-script scene images. They developed a dataset of railway station signboards that are multi-scripted. They considered the scripts which are mainly used in India like Devanagari, Gurumukhi, Bangla, Urdu, Roman, Odia, Urdu, and Telugu.

Fasil et al. (2017) discussed the script identification method using texture features like Gabor, log-Gabor, and wavelet to identify the scripts of signboards in the bus. They considered Kannada, Malayalam, and Roman scripts. In 2018 Ghosh et al. (2018) suggested the character-level script identification of natural scene text by extracting texture-based and shape-based features. In pre-processing, they used Otsu's binarization method which causes a few letters to disjoint. They overcame this problem by introducing an isotropic dilation technique. They experimented with their developed dataset which comprises Roman, Devanagari, and Bangla alphabets. In 2019 Ghosh et al. (2019a) extracted texture and topology-based features from multi-character artistic scripts. In another work in 2020 Ghosh et al. (2020), they used an extreme learning-based classifier to identify the scripts at the character level. The reported results of these works are presented in Table 7.

4.2.2.3 Text recognition It is the last stage of scene text understanding where the image-text is converted into normal text. Neumann and Matas (2013) presented a strategy that blends the benefits of sliding-window and connected component techniques. Letters were recognized in image patches which include certain strokes in a particular proportional location and with the strokes by combining the image gradient field with a series of angled bar masks. To recognize any letter geometry, a component-based tree-structured framework (Shi et al. 2013; Yildirim et al. 2013) was used to detect and recognize the letters at the same time. The suggested model was derived dynamically from the training set

Table 7 The methods used and results reported script identification using handcrafted feature-based methods in Phase 2

Author	Dataset	Technique	Performance
Singh et al. (2016)	ILST	Mid-level features	A-88.67
Singh et al. (2016)	ILST,CVSI-15	SIFT, entropy based	A-96.70
Verma et al. (2017)	500 scripts of Roman, Devanagari, Odia, Urdu, Telugu	LBP,CSLBP, DLEP	A-84.00
Fasil et al. (2017)	600 images of English, Kannada, Malayalam	Gabor, Log- Gabor, wavelet	F-0.975
Ghosh et al. (2018)	Roman-439, Devanagari-288, Bangla-306	Isotropic dilation	A-93.18
Ghosh et al. (2019a)	Bangla-1620, Roman-864, Devnagari-144	Gabor, topology based	A-93.90
Ghosh et al. (2020)	Bangla-1597, Roman-859, Devanagari-172	Gabor wavelet, GLCM, ELM	A-97.95

and represented letter components having disparate dimensions. This model was primarily motivated by the resurgence of autonomous mid-level representative development for recognition (Lim et al. 2013; Wang et al. 2013). Lee et al. (2014) argued that HOG and component-based techniques do not cover all sorts of distinct features for images with complex perspectives, and a lot of variance in typefaces. They presented a discriminative attributes pooling technique that dynamically considers the most relevant sub-regions of every scene character inside a multi-class categorization architecture, with each sub-region smoothly integrating a collection of low-level characteristics in the entire image. Yao et al. (2014b) developed strokelets, a new approach for retaining the fundamental components of letters at various granularity levels that is dynamically trained using enclosing container tags. The strokelets approach precisely recognizes distinct letters and facilitates a histogram characteristic that can be used to efficiently characterize individuals in natural scenarios. In another work Yao et al. (2014a) proposed a dictionary search algorithm for text recognition. In Fig. 9 the schematic view of their unified framework for text detection and recognition is presented.

For text recognition of natural scenes and video scripts, Raghunandan et al. (2018) presented a distinct approach where they investigated convex and concave deficits to find a candidate level. They termed this method as iterative nearest neighbor symmetry (INNS). They also introduced a concept of Mutual Nearest Neighbor Pair (MNNP) element detection to find the constituents of texts relying upon the external slope orientation of constituents. They considered the rotation connection of high and fused wavelets for character recognition. They recognize text both at the word and character levels. The text recognition results of different works are shown in Table 8.

4.2.2.4 Classifiers After features are extracted the classification is needed to categorize images or the elements of the images into the corresponding class. Verma et al. (2017) classified the scripts using K-NN (nearest neighbor) and SVM classifiers. Singh et al. (2016) classified the scripts using the SVM classification method. Fasil et al. (2017) used two types of classifiers: K-NN and SVM, to test the performance of their system. In the K-NN classifier, categorization was accomplished by providing the labeling for the highest frequency within the k training data closest to the unlabeled region which was as determined by Euclidean distance. For SVM, they employed two kernel operations: linear and Gaussian RBF. Ghosh et al. (2018) considered Random Forest, Majority Voting, Simple Logistic, and MLP (Multi-Layer Perceptron) for character-level multi-script identification. In another work, Ghosh et al. (2019a) used SVM, RBF, Random Forest, and MLP for artistic scene image analysis. They also used Ghosh et al. (2020) Bayesnet, Naive Bayes, RBF, and ELM (extreme learning machine) classifiers. Boureau et al. (2010) used linear SVM, and kernel SVM in their work whereas Singh et al. (2016), Raghunandan et al. (2018), Yao et al. (2014b) and Lee et al. (2014) used linear SVM in their work. Neumann and Matas (2013) used a nearest-neighbor classifier for character renderings depending on strokes to pick the letters from an optimally produced collection of focused areas.

Table 8 Text recognition methods and corresponding results using handcrafted features in the 2nd phase are reported

Author	Dataset	Technique	Performance
Yao et al. (2014b)	IIIT 5K-Word,ICDAR 2003, SVT	Bag of Strokelets, HOG	A-80.20, 80.33, 75.89
Raghunandan et al. (2018)	ICDAR 2013, ICDAR 2011, SVT, ICDAR 2011 BD	INNS, MNNP	W-51.60, C-61.60; W-59.60, C-68.20; W-51.70, C- 62.30; W-54.40, C-61.40
Neumann and Matas (2013)	ICDAR 2011	INNS, MNNP	R-45.40, P-44.80, F- 45.20
Yildirim et al. (2013)	ICDAR 2003	Cross-Scale Binary Features	A-85.7
Shi et al. (2013)	ICDAR 2003, ICDAR 2011SVT	Conditional Random Field model	A-79.30, 82.87, 73.51
Lee et al. (2014)	ICDAR 2003, ICDAR 2011SVT, Chars74K-15	Discriminative attribute pooling	A-0.76, 0.77, 0.80, 0.74

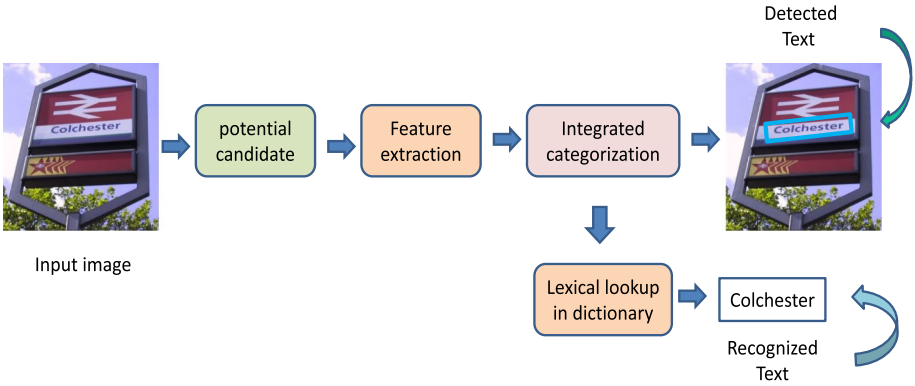

Fig. 9 A diagrammatic representation of text detection and recognition framework (Yao et al. 2014a) based on SWT and clustering

4.3 Deep learning-based techniques

Deep learning frameworks underpin almost all contemporary methodologies. Most significantly, deep learning relieves scholars from the grueling task of continually inventing and evaluating handcrafted functionalities, allowing a slew of new ideas to emerge.

4.3.1 Phase I (2012–2016)

Deep learning-based text detection, script identification, and recognition systems have emerged as significant advancements in vision-based tasks. Deep structures are made up of numerous hidden units that retrieve numerous potentially strong properties from an input image and produce the required result usually autonomously. In the very first phase (i.e., 2012–2016) of deep learning-based scene text analysis, it was seen that mainly the basic deep learning frameworks were discussed in the literature since the concept of this type of learning was new to the researchers. Different deep learning frameworks like AlexNet (Krizhevsky et al. 2012), VGGNet (Simonyan and Zisserman 2014), GoogleNet (Szegedy et al. 2015), ResNet (He et al. 2016a) etc., were published in this phase.

4.3.1.1 Text detection Scene text detection in natural images is a strong perceptual challenge. Using a collection of moderate procedures or manually constructed characteristics the issue couldn't be not entirely tackled. In 2012 Sermanet et al. (2012) devised a CNN (Convolutional Neural Network) where multi-stage features and LP pooling method (Yang et al. 2009) for real-world house number digit classification were considered. Experimenting on the SVHN Dataset they got an accuracy of 94.85%.

It was observed that MSER-based algorithms are capable of identifying the majority of textual parts within images (Neumann and Matas 2012). Nevertheless, it simultaneously produced a huge amount of non-textual items, resulting in a high level of confusion in MSERs elements of text and non-text. Furthermore, because MESR's approaches are dependent on pixel-level processes, they are extremely susceptible to noise and pixel contamination which results in erroneous interactions among letters. To solve the MSREs' multi-letter connectivity issue, Huang et al. (2014) included a CNN-based classification

Table 9 Text detection techniques and corresponding results using deep Learning-based approaches as presented in phase 1

Author	Dataset	Technique	Performance
Sermanet et al. (2012)	SVHN	CNN	A-94.85
Jaderberg et al. (2014b)	ICDAR 2003, SVT	CNN	A-91.00, 80.00
Huang et al. (2014)	ICDAR 2005, ICDAR 2011	Sliding window combined with CNN	F-0.75, 0.78
Risnumawan et al. (2016)	ICDAR 2003	CNN	A-97.40
Zhang et al. (2015)	Text-7302, non-8000	CNN, MSER, BoW	P-0.879, R-0.908, F-0.892; P-0.901, R-0.812, F-0.854
Jaderberg et al. (2016)	ICDAR 2003, SVT, ICDAR 2011, ICDAR 2013	CNN	F-91.00, 82.00, 77.00, 77.00
Yao et al. (2016)	ICDAR 2013, ICDAR 2015, MSRA-TD500, COCO-TEXT	FCN	F-0.843, 0.647, 0.759, 0.333
He et al. (2016b)	ICDAR 2005, ICDAR 2011, ICDAR 2013, MSRA-TD500	Text-CNN	F-0.79, 0.82, 0.82, 0.69

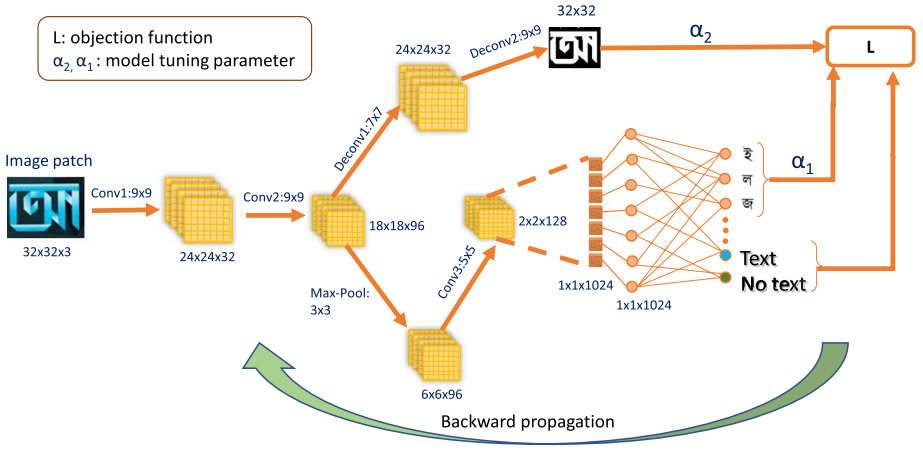

Fig. 10 The layout of the Text-CNN framework as illustrated in He et al. (2016b)

using the sliding-window and NMS approaches. This process helped in restoring the absent letters.

Researchers (Jaderberg et al. 2014b, 2016) developed a sliding window-based approach to detect text in natural images. To get rid of character segmentation/semantic segmentation or to find the text components locally (Epshtein et al. 2010; Jaderberg et al. 2014b; Neumann and Matas 2010; Yao et al. 2012) which often leads to being excluded text information, holistic approaches were used by researchers (Yao et al. 2016) where the text as a whole was considered. Based on this approach, Yao et al. (2016) estimated the likelihood of letters, text sections, and orientations of surrounding texts in a coherent structure. Using BoW (Bag of Words), MSER, and CNN methods, Zhang et al. (2015) distinguished text and non-text scene images. They evaluated the system by using their developed scene image dataset. Risnumawan et al. (2016) proposed a system that uses CNN to handle text detection challenges in low-resolution photos by considering the features from the mutual interaction among different convolutional levels of the network. Intriguingly, a level includes numerous aspects such as the incorporation of non-linear processing and maximum or average pooling. He et al. (2016b) introduced a CNN framework that concentrates on obtaining text-related areas and characteristics of image elements. This framework was trained using multi-level and rich supervised input, such as text region mask, letter labeling, and binary text/non-text information. In Fig. 10 a depiction of the structure of the Text-CNN framework is presented.

The text detection results and different techniques used in the literature are presented in Table 9.

4.3.1.2 Script identification Script identification for multilingual documents in a wild environment is a challenge. The usage of two or more scripts is extremely prevalent in multilingual and multi-script nations. As a result, in interpreting the text in these kinds of images, the determination of the underlying script is mandatory.

Shi et al. (2015) in 2015, proposed a word or line-level script identification in scene images. They experimented with the impact of multiple-stage pooling in their convolution framework with a trial-error process on different network versions by eliminating

Table 10 The techniques and results using deep learning-based script identification and text recognition as described in phase 1

Author	Dataset	Technique	Performance
Wang et al. (2012)	ICDAR 2003, SVT	CNN	F-0.64, 0.46
Goodfellow et al. (2013a)	SVHN, internal street view data	DistBelief, DNN	A-96.03, 91.00
Goodfellow et al. (2013b)	MNIST, CIFAR-10, SVHN	Maxout framework	A-65, 90.62, 97.53
Su and Lu (2014)	ICDAR2003, ICDAR2011, SVT	HOG,RNN	A-82, 83, 83
Shi et al. (2015)	SIW-10	Multi-scale pooling	A-94.4
Shi et al. (2016a)	SIW-13, CVSI-15	DiscCNN	A-89.00, 96.10
Gomez and Karatzas (2016)	CVSI-15, MLe2e	CNN, NBNN	A-96.42, 91.12
Mei et al. (2016)	SIW-13, CVSI2015	CNN, RNN	A-92.75, 94.20
Shi et al. (2016c)	IIIT5K, SVT, ICDAR 2003, ICDAR 2013	STN, SRN	A-96.20, 95.50, 98.30, 88.6

portions of the pooling tiers. Apart from using the public dataset, they prepared a dataset for their experiment. In 2016, Shi et al. (2016a) developed a deep learning-based script identification method. They used a pre-trained CNN model to extract local features and by exploiting the discriminative clustering technique, the patterns of different classes were identified from the features. These discriminative features as well as the features obtained from the CNN network were optimized and fed into another CNN framework for classification. They experimented with the SIW-13 and CVSI-2015 public datasets. Gomez and Karatzas (2016) considered CNN-based features and Naive-Bayes Nearest Neighbor classifier(NBNN) fine-grained categorization properties. To consider features, as well as spatial dependencies in the textual parts Mei et al. (2016) proposed a script identification technique using CNN and RNN (Recurrent Neural Network). They adopted an average pooling structure to cope with the arbitrary image dimension. Across word vectors, this method makes use of image interpretation and spatial relationships. Whereas CNN creates detailed picture understandings and RNN successfully investigates long-term spatial relationships, the technique integrates CNN and RNN under a single end-to-end trainable system. They tested the proposed method using publicly available datasets: SIW-13 and CVSI2015. The techniques for script identification and text recognition results are summarized in Table 10.

4.3.1.3 Text recognition Several real-world systems, including navigation, self-driving cars, and graphics interpretation, are made possible by recognizing text in photos. Text recognition in natural images (Jaderberg et al. 2014a; Goodfellow et al. 2013a, b; Shi et al. 2016c; Su and Lu 2014; Wang et al. 2012) has piqued society's attention for such objectives. Wang et al. (2012) expressed the issue of end-to-end text recognition by dividing the task into two parts: text localization and word recognition. In text localization, they identified individual words or lines of text. For the recognition of text, they designed a CPCPD (convolution-pooling-convolution-pooling-dense) architecture. They used ICDAR 2003 and SVT datasets to test the performance of the system. Text recognition in an unrestricted/wild environment is a tough issue that has sparked increased scientific attention in subsequent years. There were numerous approaches proposed to overcome this issue. Goodfellow et al. (2013a) demonstrated an integrated deep framework to recognize multi-digit numerals from roadside images. A framework named Maxout network (Goodfellow et al. 2013b) which enhances the performance of the network and at the same time minimizes the dropout was developed. Su and Lu (2014) extracted features by applying HOG features. The RNN and CTC (connectionist temporal classification) algorithms were adapted for text recognition. Often, viewpoint deformation causes aberrant geometries and curled/bent letter placement in scene images. To tackle such cases, Shi et al. (2016c) designed a framework which is based on Spatial Transformer Network (STN) and a Sequence Recognition Network (SRN) where the STN is responsible for text irregularity correction and SRN is deployed for text recognition.

4.3.2 Phase-II (2017–2021)

In this phase, the researchers kept their attention more on complex scene images. Different advanced architecture (Cheng et al. 2017; Dargan et al. 2020; Lyu et al. 2018a; Shi et al. 2018; Zhan and Lu 2019) were developed to handle the issues related to text string identification and recognition in challenging natural images.

The majority of current research has focused on enhancing scene text recognition efficiency by incorporating additional resilient and functional graphic characteristics, like strengthening backbone systems (Cheng et al. 2017; Lyu et al. 2018a), introducing reconfiguration components (Shi et al. 2018; Zhan and Lu 2019) and enhancing attention processes (Ma et al. 2021a; Huang et al. 2019; Wojna et al. 2017; Yang et al. 2017). Nonetheless, it's indeed true that a person's comprehension of scene text is influenced not just by sensory perceptions knowledge, but by the elevated textual linguistic contextual interpretation. The advanced deep-learning models like Refinenet, YOLO-V1 (You Only Look Once), YOLO-V2, PSPNet, Fast-FCN (Fully Convolutional Networks), and DRN (Dilated Residual Network) (Lin et al. 2017; Redmon and Farhadi 2017, 2018; Wang et al. 2018; Wu et al. 2019a; Yu et al. 2017; Zhao et al. 2017) etc., are being proposed in this phase.

4.3.2.1 Text detection The presence of text in natural scene images is important because by knowing apriori the computational cost can be reduced by not processing non-text images.

Bai et al. (2017) discussed the block-level classification of text and non-text scene images. They designed a multi-scale Spatial Partition Network (MSP-Net) for classification purposes. In Sriman and Schomaker (2019) authors discussed texture feature-based methods (color autocorrelation histogram and scale-invariant feature transform) to extract color features to identify the text/non-text scene images. They used 1-NN and SVM for classification.

Ghosh et al. (2019b) in 2020 proposed a six-layered CNN framework to segregate the text/non-text in natural images. They used MSRA-TD500, ICDAR 2003, and SVT as text scene image datasets and employed the 15-Scene Image Dataset as the non-text dataset. A new feature descriptor using skeletonization and distance transform process was proposed in Khan and Mollah (2019). They developed a dataset that comprises Bangla, Devanagari, and Latin scripts. For classification, a CNN-based framework was designed.

Text detection in scene images has attracted growing interest from the field of computerized imaging and offers a variety of applicability in content processing, robot control, traffic navigation, OCR interpretation, information extraction, virtual reality, etc. It is indeed an intractable concern due to a wide range of text variations in colors, fonts, directions, languages, and dimensions, and also highly complicated and text-like backgrounds, and several obfuscations and artifacts induced by image captures such as inconsistency lighting, poor contrast, blurriness, and obstruction, etc. Considering the incredible advancement of deep learning-based techniques, many CNN/RNN premised object detection structures were developed to fix these issues which include Faster R-CNN (Region-Based Convolutional Neural Network), attention-based frameworks, FCN, etc., which significantly outclasses conventional MSER, SWT oriented detection strategies.

Some advanced deep learning-based approaches like mask R-CNN, graph-based, attention network-based, etc., are also adopted for text detection and are described in the later half of this section.

Zhou et al. (2017) proposed a scene text detector named EAST (Efficient and Accurate Scene Text Detector) using VGG-16 (Simonyan and Zisserman 2014) and PVANET (Kim et al. 2016) models. To get the correct bounding box among the closely occurring bounding boxes, they developed a local non-maximum suppression technique. Using the public datasets (Yao et al. 2012; Karatzas et al. 2015; Veit et al. 2016), they showed the frame per second (FPS) value improved over the state-of-the-art. In 2019, a deep neural network-based framework was developed by Ghosh et al. (2019c) using a pre-trained EAST model to localize the text in natural images. Zhang et al. (2018) proposed a feature enhancement

network for text box selection and by using an adaptively weighted position-sensitive region of interest (ROI) the positions of the boundary boxes were fine-tuned. They used a positive mining technique for dataset equalization. They experimented with ICDAR 2011 and ICDAR 2013 datasets and achieved 0.896 and 0.941 precision respectively. The techniques of extraction of arbitrarily shaped texts in scene images were proposed by Xu et al. (2019a). By considering the orientation of the text and using the proper mask they created a model. Their main challenge was to localize the curved or rounded-shaped text. The experimentation was done using the public datasets (Ch'ng and Chan 2017; Karatzas et al. 2015; Gupta et al. 2016; Yuliang et al. 2017; Yao et al. 2012). Khalil et al. (2021) suggested a fusion-based strategy by using Resnet-50 and EAST models to improve the EAST model's performance. Using the ICDAR 2017 MLT dataset the F-score increased by 0.91% in their approach. It was observed that the performance of the EAST model drops for oriented/inclined and curved text segments in images. Ghosh et al. (2021b) proposed an M-EAST model to overcome the issues in the EAST model and got an improved FPS (frame/second) value which is 4.17 times higher compared to EAST.

Zhu and Du (2021) conferred a method of text detection in scene images by the concept of the "mountain". They treated the text center as a mountain top and the text border as a mountain bottom. They proposed text center-border probability (TCBP) and text center-direction (TCD) methods to identify the top and bottom of mountains. They showed their method can deal with oriented and curved text as well. They tested their system with ICDAR2015, SCUT-CTW1500, RCTW-17, and MLT datasets. To identify randomly oriented texts in the scene image, Wang et al. (2021a) suggested the Rotational You Only Look Once (R-YOLO) which is a CNN-based framework. A rotating anchor box in different directions for enclosing text bounding boxes was used. The attributes of different scales were retrieved to evaluate the text's likelihood and oriented bounding boxes. Non-maximum suppression along with rotational distance intersection techniques to minimize replication and to have proper boxes was used. ICDAR2013, ICDAR2015, ICDAR2017, MLT, and MSRA-TD500 datasets were deployed to test the efficiency of this approach.

Pandey et al. (2021) proposed a text extraction and recognition method in scene images considering deep neural network (DNN) based techniques. To consider only the text scene images, they used a weighted naıve Bayes classifier (WNBC) for text and non-text classification. For text extraction, the DNN-based adaptive galactic swarm optimization (AGSO) technique was used. Character recognition was performed by DNN and AGSO methods. They also suggested an algorithm to minimize the errors generated in text/non-text classification. The whole methodology was tested using the IIIT5K dataset. An end-to-end text localization as well as clustering of scripts was proposed by Munjal et al. (2021) in 2021. They proposed the OnDevice Text Localization with Clustering of Script (TeLCoS) technique for localization and clustering texts. The weights in the network were tuned by using shared convolution units. With the help of a channel pruning strategy, this could be used in low-resourced devices. ICDAR-2013, ICDAR- 2017, and MSRA-TD500 datasets were used in their experimentation.

Liu et al. (2021) used an efficient OBD (Orderless Box Discretization) tool to identify textual content in multi-orientation scenes. OBD can overcome the inaccurate marking problem by modeling the pointwise estimation into orderless edges using discretization techniques. They suggested a basic yet efficient matching-type learning approach to recreating the quadrilateral bounding box to decipher correct vertex representations. To investigate the theoretical maximum bound of their system, they performed detailed embolization analyses on a few training elements: data organization, pre-processing, framework creation, concept development, scores generated from different predictions. They applied

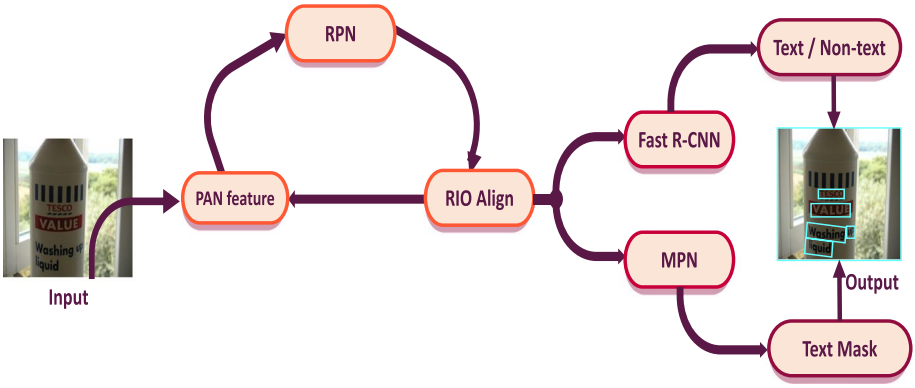

Fig. 11 The structure of the Mask R-CNN driven text detector model (Huang et al. 2019) based on RPN, PAN, Fast R-CNN, and MPN

their methods to ReCTS, ICDAR 2015 Incidental Scene Text, and The ICDAR 2017 MLT datasets for scene text identification.

To achieve higher-level graphical features and increase text identification and recognition performance, Naiemi et al. (2021) created a CNN-based framework. In this analysis, a pre-trained ResNet-50 network was utilized to obtain low-level graphical features. In their proposed framework, an upgraded ReLU layer module with a specified responsive system with a wide potential to perceive text elements had the potential to distinguish text elements even on curved landscapes. We also introduced a character detection pipeline architecture that is resilient to unusual (curve and vertical) text. They proposed the local word directional pattern (LWDP) method to encode pixel values that emphasize the texture of the characters. The testing was done by leveraging ICDAR 2013, ICDAR 2015, and ICDAR 2019 datasets.

Gkioxari et al. (2015) proposed an R-CNN framework to predict multiple areas of localization. Improving this framework to use in scene text analysis, faster R-CNN was developed by Girshick (2015). He et al. (2017) in 2017 developed the Mask R-CNN framework by extending the concept of faster R-CNN. It introduces an object mask prediction stream in tandem with the current bounding box recognition stream. It leveraged ResneXt as the basic network and developed RoIAlign (He et al. 2017) to replace RoIPool (Girshick 2015) to correct the pixel alignment. Mask R-CNN focused text identification method that detects multi-oriented and bent text in real scene images in a coherent way. In 2019 Huang et al. (2019) suggested a framework to improve Mask R-CNN's feature rendering abilities using pyramid attention network (PAN) and region proposal network (RPN), faster R-CNN, and mask predictor network (MPN). In Fig. 11 the architecture of their proposed framework is shown. Leveraging the conventional object identification model, Mask R-CNN, Ammirato et al. (2019) constructed a stochastic method for object identification.

The architectural information inside the data could be incorporated to characterize the relationships between objects and provide greater potential understanding underneath the data by expressing this as graphs. The combination of deep learning and graph-based architectures (Gao et al. 2020; Liu et al. 2018a, 2020; Mafla et al. 2021; Ma et al. 2021a; Shi et al. 2017a) are becoming popular for the cases of text clustering to generate semantic textual information.

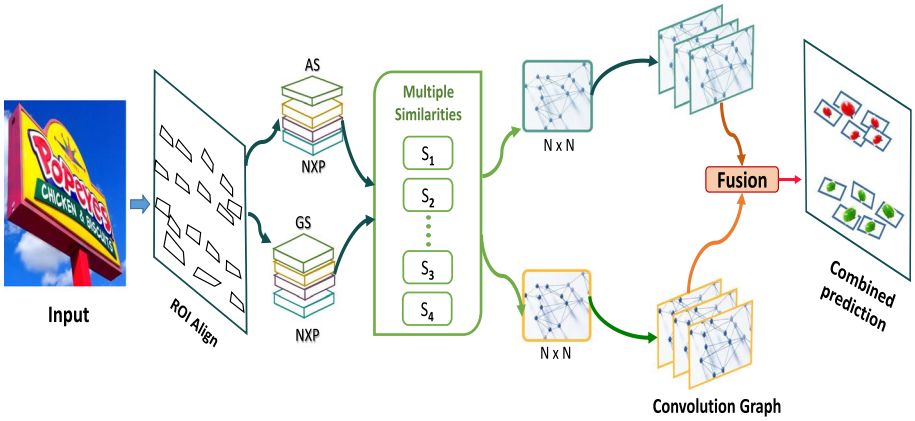

Fig. 12 The MSGCN framework as proposed by Liu et al. (2020) in 2020

Shi et al. (2017a) estimated both object sections and linkages among multi-resolution attribute mappings, based on the SSD (single shot detector) (Liu et al. 2016a) technique. Bounding boxes at the individual stages were created by combining inclined rectangles based on their connection ratings. During scene text identification, Liu et al. (2018a) in 2018 introduced a robust Markov Clustering Network (MCN). They represented object identification as a graph-based grouping issue and constructed a fully adaptable framework using a stochastic flow graph, enabling dynamic scene text detection. A graph convolutional network (GCN) for semi-supervised classification was proposed by Kipf and Welling (2016) in 2016. For ordering the text sections one after another, linking with the individual areas are necessary. To overcome the difficult text-line clustering situation, Ma et al. (2021a) proposed a GCN-based perceptual association identification architecture to offer a novel arbitrarily structured text spotting methodology. To handle textual cases with large inter-character or very short inter-line spacing, the perceptual connection assertion of the text-line combining strategy was adopted. According to them, the proposed GCN can substantially manipulate contextual knowledge to increase link predictive performance. They used MSRA-TD500, RCTW-17, CTW1500, DAST1500, and Total-Text datasets to test the performance of their approach. Authors in Liu et al. (2020) considered the text sections as contextual graphs, with every vertex representing a section. Edges reflect segment-to-segment relationships. To perform inference on sectional contextual graphs from both visual and shape standpoints, they used the GCN-based similarities among text sections and termed their network as multiple-similarity GCN(MSGCN). In Fig. 12 the MSGCN architecture is shown. Gao et al. (2020) designed a multi-modal GCN for a visual question-answering system on scene text images.

To take advantage of the complementing data given by the systems, Mafla et al. (2021) integrated context clues to conspicuous image areas. By understanding a shared conceptual area among conspicuous items and text detected in the image, a GCN-based framework was proposed which was based on the work of Kipf and Welling (2016) to conduct inference. GCN and its variants (Li et al. 2019b; Chen et al. 2018; Ma et al. 2021a; Wu et al. 2019b) were used in different scene text image analysis problems. Combining GCN and RNN, an adaptive boundary proposal framework was developed by Zhang et al. (2021a) in 2021.

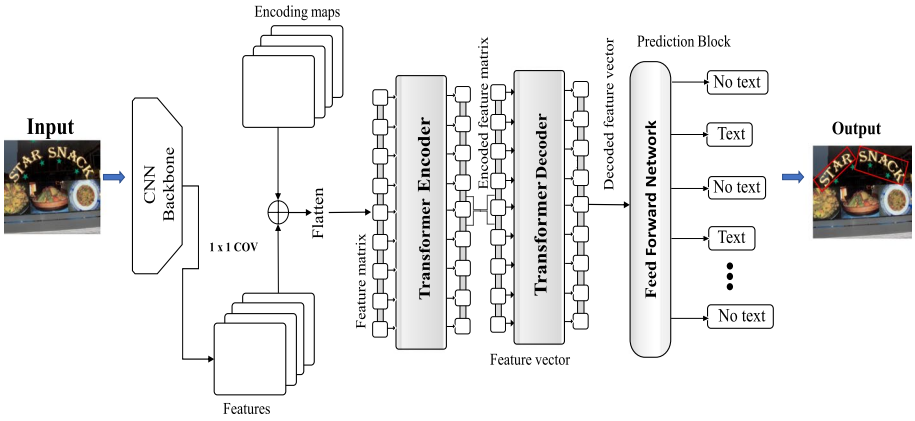

Fig. 13 A transformer-based text detection approach as presented in Raisi et al. (2021)

The traditional encoder-decoder framework is usually incapable of effectively analyzing lengthy input strings. Since only the final hidden layer of the encoder is utilized as the background vector for the decoder. Encoders cannot find the actual location of the characters since it deals with spatial data. The Attention Mechanism, on the other hand, explicitly solves this problem by retaining and using all hidden states of the input series during the decoding stage. It was accomplished by making a one-of-a-kind representation across each time phase of the decoder performance and all of the encoder's hidden states. This ensures that the decoder has exposure to the whole input sequence with every output and can specifically take out unique components to generate the result.

To avoid false text-like structures in the image, Huang et al. (2019) used the mask R-CNN framework for working with scene images. They proposed a pyramid attention network that worked with the R-CNN network to enhance features for curved and multi-oriented text detection in natural images. They experimented with ICDAR-2015, ICDAR-2017 as multi-oriented, and SCUT-CTW1500 as curved text image datasets. Cao et al. (2020) recommended a multi-oriented text identification algorithm in conjunction with an attention component. This approach detects text using pixel-by-pixel prediction and is premised upon FCOS (fully convolutional one-stage object detector) system (Tian et al. 2019). In 2021 Qin et al. (2020) designed a soft-attention mechanism for scene text detection. The performance measures using Precision, recall, and f-score of the state-of-the-art deep learning-based methods experimented on various datasets for scene text detection are tabulated in Table 11.

In 2017 a research team of Google presented a new framework named Transformer (Vaswani et al. 2017) to replace networks based on RNN. The transformer is a deep learning framework that uses the self-attention process and component-wise weighting scheme. Raisi et al. (2021) in 2021 proposed a Transformer-based framework (shown in Fig. 13) which is based on an attention-based mechanism that is able to handle and detect the multi-oriented, curved, irregular shape text bodies in natural images. The encoder's self-attention stack is able to distinguish distinct words in the scene image While the decoder pays attention to distinct letters in words by employing various trainable feature vectors/ object queries.

4.3.2.2 Script identification Recent end-to-end scene text recognition algorithms (Fang et al. 2018; Li et al. 2019a; Wang et al. 2021b) presumed a singular script supplied in advance. The topic of unrestricted text interpretation for massive sets of images from unverified senders needs to be addressed. For this, a substantial study was done on script identification even in complicated scenarios (Cheng et al. 2019; Lu et al. 2019; Ghosh et al. 2021b; Zdenek and Nakayama 2017).

According to Gomez et al. (2017) state-of-the-art overlooked the crucial feature of scene text situations due to its changeable aspect ratio. They proposed a patch-based identification scheme to maintain exclusionary elements of the picture that are distinctive of its category. Rather than scaling input images to a predetermined aspect ratio as is usual in the usage of comprehensive CNN models were developed for this scheme.

To discriminate images, stroke-part generation was considered by using ensembles of conjoined frameworks. Gomez and Karatzas (2016) developed a new public baseline dataset named MLe2e (Gomez et al. 2017) for evaluating all phases of end-to-end visual comprehending text. Zdenek and Nakayama (2017) used the local convolutional information in conjunction with the classic bag-of-visual-words method and presented a technique for script identification in scene images. The classification was done using SVM and MLP techniques. The performance of the system was tested on benchmark datasets like SWI-13, MLe2e, and CVSI2015.

Tounsi et al. (2017) compared using two pre-trained CNN models AlexNet and VGG-16 which were trained from scratch, to present the scene image identification. The NBNN classifier's fine-grained categorization properties were combined with the emotive description of convolutional information with this technique. A novel accessible standard dataset was also developed for the assessment of all phases of end-to-end scene text interpretation techniques. Cheng et al. (2019) proposed a patch-based method where the patches in images are fed to convolutional layers for feature extraction. Initially, features were extracted by a convolutional layer which was followed by a patch aggregator and a squeezer module to have local and discriminative features from the patches. A weighted fusion strategy was adopted to fuse the prediction of the two modules. They tested the performance of their proposed approach using SIW-13, RRC-MLT 2017, and CVSI 2015 datasets.

Lu et al. (2019) proposed a method of script identification in natural images by integrating local and global features from CNN frameworks which were developed on the foundation of ResNet-20 network architectures. The Adaboost method was taken for conclusive level fusion obtained from the CNN.

They used the image as well as the video script dataset to test the efficiency of their system. SIW-13, CVSI-2015, MLe2e, and ICDAR-2017 were deployed in the experimentation. Attention-based patch weights technique was considered by Bhunia et al. (2019) for script identification in natural frames. They used the CNN-LSTM (Long Short term Memory) method to extract global features. Local features were extracted from the divided parts of the images after CNN was applied and considering the weights from the attention-based patches. Using a dynamic fusion strategy, the local and global features were united for individual image parts. They experimented with SIW-13, ICDAR-17, MLe2e images datasets, and the CVSI-15 video scripts dataset. Ghosh et al. (2021a) developed a lightweight CNN framework for video script identification. The performance of the system was tested on the CVSI-15 dataset in noiseless and disparate noisy scenarios. Ma et al. (2021b) emphasized on feature extraction and classification for script identification in scene images. They used CNN based framework for feature extraction and used a residual attention model for considering features from characters that had background and foreground similarity. The

classification was accomplished by another CNN framework which used a global pooling layer. They experimented with SIW-13, RRC-MLT2017, MLe2e, and CVSI-2015 datasets.

The natural scene images consist of simple/complex backgrounds. But, in the case of movie poster images, the complexity becomes very high. Apart from complex background design/graffiti, there are different texts available in disparate sizes, colors, textures, and orientations seen in movie poster images. Ghosh et al. (2022) suggested a deep learning-based framework to identify the scripts of the movie titles. In their other work (Ghosh et al. 2021b) in 2021, they proposed a shallow convolution neural network (SCNN) for movie title identification. In Table 12 the state-of-the-art techniques for script identification and results are summarized.

4.3.2.3 Text recognition Deep learning-based autonomous systems demonstrated higher efficiency (Bai et al. 2018; Cheng et al. 2017; Gao et al. 2019; Li et al. 2019a; Shi et al. 2018; Wang et al. 2020b; Yu et al. 2020) in text recognition in natural images. Certain systems recognize text at the letter scale (Bai et al. 2018; Kong et al. 2019), whereas the majority of techniques recognize text at the word/sentence scale (Cheng et al. 2018; Lyu et al. 2018a). The latter is chosen because the annotating process is simpler and less time-consuming. Dizaji et al. (2018) developed HashGAN which is a deep unsupervised hashing technique that effectively extracts binary representations of input images.

There are other advanced techniques based on GAN, end-to-end, ensemble-based attention, attention-based Encoder-decoder, sequence-to-sequence attention model-based, etc., that were also discussed in the literature and are presented in the later half of this section.

Several attention-based approaches (Cheng et al. 2017; Kim et al. 2017; Li et al. 2019a), majorly struggled with alignment difficulty as a consequence of their orientation procedure, which depends upon previous decoder outcomes. Wang et al. (2020b) addressed this issue and proposed a decoupled attention network (DAN), that separates the alignment process from making use of previous decoder findings. This network serves as an end-to-end text recognizer that is efficient, adaptable, and resilient.

To effectively retain semantic features of text, RNN-based architectures (Lei et al. 2018; Liu et al. 2018b) were investigated. Nevertheless, RNN-based approaches have had certain evident flaws, including the time-dependent interpretation and one-end sequential temporal contextual propagation, that effectively restrict the use of meaning and computing effectiveness. To tackle these issues, Yu et al. (2020) offered an innovative end-to-end trainable system for efficient scene text recognition called the semantic reasoning network (SRN), which includes a global semantic reasoning module (GSRM) that captures universal semantics context via multi-way concurrent propagation. Sajid et al. (2021) presented a text recognition system to deal with the difficulties in scene-text recognition. They proposed a multi-scale and scale-wise spatially monitored network to retrieve multi-scale features. They argued that along with feature extraction, this network can perform spatial attention at the same time. In a more useful method, distinct feature sizes may be expressed clearly. Also, their system experiences multi-scale integration with one another. They adopted an approach of interscale fusion in this work. They applied their methodology on SVT-P, CUTE80, IIIT-5k, SVT, ICDAR 2003, ICDAR 2013, and ICDAR 2015 datasets.

In 2014, Goodfellow et al. (2014) presented the generative adversarial network (GAN) as a deep learning framework. It depicted generative designing as a contest involving two systems: a generator system generates data in the presence of noise, while a discriminator system separates the generator's output from actual information. The generator can provide quite excellent performance with some training. Xu et al. (2019b) proposed an approach

based on a generative adversarial network to get rid of the dealings of foreground and background color and extraction of text using the connected component method. They experimented on KAIST and MSRA TD 500 datasets. Kong et al. (2019) developed a generative adversarial recognition network (GARN) for scene character recognition. The designed architecture replaces the homogeneous variation and discriminator in GANs with a Gaussian mixed distribution and multinomial predictor.

Word detection and recognition are closely associated activities. They may share feature details. Furthermore, both jobs can be used in tandem. Correct identification of text sections aids in good recognition performance. The recognition results could well be utilized to fine-tune the accuracy of text detection. In Li et al. (2017) proposed an end-to-end method for the detection as well as recognition of text in real-world images which are depicted in Fig. 14. They detected the region of interest using a text proposal network (TPN), and TDN (Text Detection Network). Using the Text recognition module and attention mechanism the words are recognized. But their method was not able to handle images having oriented text regions. In 2019 by replacing the 1D attention mechanism with 2D attention and considering TPN and attention-based technique, they Li et al. (2019a) extracted character-level features. They argued that their proposed attention technique was tuned to take into account specific features, which improves recognition efficiency. Yang et al. (2019) presented a symmetry-constrained rectification network (ScRN) to integrate existent detection systems for developing a single architecture from start to finish. Wang et al. (2021b) concentrated their work on detection text that is oriented along with curved positions. In the end-to-end context, they estimated attention values which were used to determine the optimized oriented bounding box. Their architecture was fully trained in a straightforward end-to-end manner. During the training phase, both detection and recognition operations were combined and refined. The generated RoI featured vectors that considered the image size of various words into account and retrieved the appropriate content for further detection and identification. Khalil et al. (2021) modified the EAST detector model and along with a fully connected neural network model they designed an end-to-end model for text detection and script identification in scene images.

In ensemble-based attention models along with attention-based recognition systems which decode feature sequences, there is module(s) where localized contextual information among neighboring local attribute arrays are considered for achieving higher performance. Fang et al. (2018) showed that the performance of the system can be improved by ensembling the attention and language components together. For training the system, they incorporated numerous losses via graphical signals and verbal constraints. Gao et al. (2019) reasoned that the positions of the input image patch successions and the resulting character succession are highly correlated. Nevertheless, while identifying the present character, many modern recognition methods seldom take into account this regional detail from the input sequencing. They described a Local Restricted Attention (LRA) method that encodes the current vector by taking neighboring vectors from the input sequence into account. They presented an ensemble decoder block that integrates the LRA and conventional decoding mechanisms. This module not only improves text recognition performance significantly while taking less time to train, but it can also be readily integrated into different recognition systems. IIIT-5K, SVT, CUTE80, SVT-Perspective and ICDAR 2003,2013,2015 datasets were used in their work. Zhang et al. (2021b) presented an ensemble of three different attention modules for text recognition.

Using an attention-based encoder-decoder structure, scene text analysis is recently practiced since it bestows good potential outcomes on numerous standard activities (Cheng et al. 2017; Li et al. 2019a). Present text recognition research has been driven by the

encoder-decoder architecture. Several platforms built using such architecture have reached cutting-edge functionality. It decodes an output label series in an auto-degenerative manner after encoding an input image as a one-dimensional characteristic series or a two-dimensional feature space, focusing on a particular portion at every sampling interval. Various encoders, decoders, and attention mechanisms were studied in the research. But it was found that this procedure confuses and misleads if there is a mismatch among the actual pattern obtained from the attention's output produced by missing or extraneous letters. To address this issue, Bai et al. (2018) in 2018 suggested an edit-likelihood technique that takes into account not only the probabilistic model but also the possibility of absent or extraneous letters. In the same year, Cheng et al. (2018) set out to tackle directed text and discovered that the existing encoder-decoder system struggles to grasp the inclined text's underlying properties. To retrieve text attributes in images along those orientations, they converted the source image into four pattern sequencing of four quadrants.

A text image is a collection of various characters that can be considered as a varying labeling sequence. The most widely used domain adaptation approaches (Pei et al. 2018; Yang et al. 2018; Zeiler et al. 2012; Zhuo et al. 2017) could not be straightway extended to sequence prediction since a generic adjusted approximation excludes critical fine-grained details at the character level, which results in inadequately characterizing the content.

In 2019 Zhang et al. (2019) proposed a network, based on the attention network, termed as Sequence-to-Sequence Domain Adaptation Network(SSDAN) for text recognition in images. They devised a gated attention similarity module to bridge the gap between the target and obtained feature space by extracting character-level features. The flowchart of their proposed method is presented in Fig. 15. In the same year, Sheng et al. (2019) pointed out the sluggishness in the training pace of intrinsic recurrence of RNN and the complexity involved in stacked convolutional networks for enduring feature retrieval. They proposed a no-recurrence sequence-to-sequence text recognizer that deals with the recurrences and convolutions altogether. Their network utilizes the encoder-decoder network, in which the encoder extracts image attributes using stacked self-attention. Using piled self-attention, the decoder recognizes texts. They claimed using a self-attention mechanism, the training was done parallel and exposed low complexity. They used a modality-transform module to adapt 2D input images into 1D patterns which were conglomerated with an encoder to retrieve additional exclusionary information from images. They used CUTE80, IIIT5K, SVT, ICDAR 2003, ICDAR 2013, ICDAR 2015, and SVT-P benchmark datasets to test the efficiency of their method. In 2021 Aberdam et al. (2021) proposed a sequence-to-sequence contrastive learning framework for scene text recognition. In this network, each feature vector is partitioned into several occurrences, which are used to estimate the discriminative penalty. This procedure allows contrast on a sub-word scale, extracting numerous affirmative pairings and many adverse instances in each image. They also proposed unique enhancement algorithms, various encoder designs, and bespoke visualization components to provide visualizations in text recognition. RIMES, IAM, CVL, Synth-Text, IIIT5K-words, ICDAR-2003, and ICDAR-2013 datasets were deployed for accuracy testing. The changes in accuracy with the effect of noise, and blurriness would make their work more interesting.

Fang et al. (2018) showed that convolutional planes can serve simultaneously as the encoder and decoder in a scene text recognition framework. The encoder uses two-dimensional convolution and the decoder uses an attention module to record graphic inputs. A linguistic component was also made to simulate verbal constraints, which can be considered an assembly for making projections. This component is constructed

on one-dimensional convolution followed by gated linear units (GLU) (Dauphin et al. 2017). Repeated attention and language failures were gathered to train the system's completion. Lyu et al. (2018a) presented a separation approach for word spotting that leverages an FCN-based technique for recognition. Liao et al. (2017) introduced an approach named character attention FCN (CA-FCN), which represented the issue in two dimensions. The system can efficiently distinguish abnormal and normal textual occurrences by conducting letter categorization at every visual point.

Raisi et al. (2020) presented a framework based on fusing the Transformer and a 2D stationary encoder. In order to retain the spatial information in 2D image features and strengthen the encoder component's ability to capture the characteristics produced by the encoder's self-attention technique, they added a separate feed-forward network tier to the encoder unit. Zhu and Zhang (2021) proposed a transfer model for end-to-end-to-scene text recognition. Atienza (2021b) designed a Transformer named Vision Transfer for Speedy and Effective recognition of both regular and irregular text. To improve the accuracy Tao et al. (2021) proposed a Transformer framework named Transformer-based text recognizer (TRIG). Compared to Atienza (2021b) their inference time to process an image is 6.6 times less.

In Table 13 different text recognition methods and corresponding results are depicted.

4.4 Results

4.4.1 Evaluation protocol

The process of text detection was assessed by employing different protocols like ICDAR 03 (Lucas et al. 2005) (considered best match among text rectangles), DetEval (Wolf and Jolion 2006)(paid attention towards many matches), IoU (Karatzas et al. 2015), Yao (Yao et al. 2012) (concentrated on arbitrary orientations), TedEval (Lee et al. 2019) (character-level detection), etc. For script identification, protocols like accuracy, precision, sensitivity, specificity, etc., Ghosh et al. (2021a) were used. The text recognition was assessed utilizing word recognition accuracy or an end-to-end recognition (Wang et al. 2012).

Results obtained by the researchers in scene text including text detection, script identification, and recognition are presented here. We have followed the same structure here as followed in section 4.2.1 and 4.2.2 including both handcrafted feature-based as well as deep learning works already reported.

4.4.2 Text detection

Yao et al. (2012) trained their system with individual ICDAR 2005 dataset along with the mixture set of their developed MSRA TD-500 and ICDAR 2005 datasets. Their system was evaluated using MSRA TD-500, ICDAR, and Oriented Scene Text Database (OSTD) with standard metrics like precision (P), recall (R), and f-score (F) and obtained 0.69, 0.66, 0.67 and 0.68, 0.66, 0.66 while tested on ICDAR and 0.63, 0.63, 0.60 and 0.53, 0.52, 0.50 using MSRA TD-500 and 0.77, 0.73, 0.74 on and 0.71 0.69 0.68 on OSTD dataset using mixture and ICDAR as the training set respectively. Similarly, the other results are detailed in Table 4 which contains results of works described in Sect. 4.2.1 the first phase of handcrafted-based works text detection is tabulated. From this table, it can be found that

Table 11 The techniques and corresponding results for text detection using deep learning-based methods as presented in the 2nd phase

Researchers	Dataset	Methods	Performance
Busta et al. (2017)	ICDAR 2013, ICDAR 2015	RPN, CNN	P-0.92, R-0.89, F-0.81; FPS-10; P-0.58, R-0.53, F-0.51, FPS-9
Zhou et al. (2017)	COCO-Text, MSRA-TD500	VGG16, PVANet, NMS	R-0.324, P-0.5039, F-0.3945; R-0.6743, P-0.8728, F-0.7608
Shi et al. (2017a)	ICDAR 2015 Incidental scene text, MSRA-TD500, ICDAR 2013	CNN based on VGG16	P-73.10, R-76.80, F-75.00; P-86.00, R-70.00, F-77.00; P-87.70, R-83.00, F-85.30
Lyu (2018 et al.)	ICDAR2013, ICDAR2015, ICDAR2017, MSRA-TD500, MLT, COCO-Text	VGG 16, Corner detection, position-sensitive segmentation	P-93.30, R-79.40, F-85.80, FPS-10.4; P-94.10, R-70.70, F-80.70, FPS-3.6; P-87.60, R-76.20, F-81.50, FPS-5.7; P-74.3, R-70.60, F-72.40; P-61.90, R-32.40, F-42.50
Liu et al. (2020)	ICDAR2013, ICDAR2015, MSRA-TD500	MSGCN	P-88, R-87, F-88; P-72, R-80, F-76; P-88, R-79, F-83
Huang et al. (2019)	ICDAR-2015, ICDAR-2017 MLT, SCUT-CTW 1500	PAN, RPN, Fast R-CNN detector	P-0.815, R-0.908, F-0.859; P-0.698, R-0.800, F-0.743;P-0.832, R-0.868, F-0.850
Xu et al. (2019a)	ICDAR2015 Incidental Scene Text, MSRA-TD500	VGG-16	P-0.843, R-0.839, F-0.841, FPS-1.8; P-0.874, R-0.759, F-0.813
Ammirato et al. (2019)	MSCOCO	Mask-RCNN	Score- 16.36
Ghosh et al. (2019c)	IC03, SVT	EAST	P-0.8313, R-0.7605, F-0.7943;P-0.6972, R-0.601, F-0.6455
Decker1a et al. (2020)	IC11, IC13	MobileNetV2, SSD	P-97.40, R-94.81, F-96.09; P-88.38, R-66.67, F-76.00
Gao et al. (2020)	IC11, IC13	GNN	P-97.40, R-94.81, F-96.09; P-88.38, R-66.67, F-76.00
Cao et al. (2020)	IC13, IC15, MSRATD-500	FCOS network	P-93.20, R-84.60, F-88.70; P-89.00, R-81.20, F-84.90; P-84.20, R-71.20, F-77.20
Qin et al. (2020)	CTW1500, IC15, Total-Text	Attention-based	P-81.80, R-76.80, F-79.40; P-82.86, R-80.20, F-81.56; P-82.28, R-76.48, F-79.30
Wang et al. (2021b)	IC13, IC15, Total-text, COCO Text	TDN	F-96.39, 87.80, 61.25, 64.43

Table 11 (continued)

Researchers	Dataset	Methods	Performance
Wang et al. (2021a)	IC13, IC15, IC17 - MLT	R-YOLO	P-82.90, R-90.10, F-86.40; FPS-47.00, P-78.20, R-87.00, F-82.30 FPS-62.50; P-71.70, R-77.10, F-74.30, FPS-67.60
Munjal et al. (2021)	IC17, IC13	TeLCoS	P-78.7, R-64.90, F-71.12; P-93.56, R-86.72, F-90.50, FPS-83.90
Liu et al. (2021)	IC15, IC19, MLT	OBD	P-88.20, R-92.10, F-90.10; P-76.44, R-82.75, F-79.47
Naiemi et al. (2021)	IC13, IC15, IC19	CNN-based framework, Upgraded ReLU, inception models	P-0.9283, R-0.9463, F-0.9372; P-0.9100, R-0.9250, F-0.9174; P-0.6908, R-0.7119, F-0.7012
Khalil et al. (2021)	IC17, MLe2e	Fusion of Resnet-50, EAST model	F-54.34, 81.13
Pandey et al. (2021)	IIIT5K	DNN, SWT, AGSO	P-93.79, R-96.8, F-95.2
Munjal et al. (2021)	IC13, IC17, MSRA TD500	TeLCoS	P-87.2, R-76.9, F-81.72
Ma et al. (2021a)	RCTW-17, MSRA-TD50, Total-Text, CTW1500, DAST1500, SynthText	GCN	P-75.90, R-61.70, F-68.10; P-90.5 R-83.2, F-86.70; P-84.80, R-83.10, F-84.00; P-86.20, R-83.30, F-84.8, FPS-10.60; P-89.00, R- 82.90, F-85.80
Ghosh et al. (2021b)	Bangla-484, Roman-607, Devanagari-340	MEAST	P-84.09, R-81.52, F-82.77, FPS-22.27
Zhu and Du (2021)	MLT, IC15, RCTW-17, SCUT-CTW1500	FPN	P-79.33, R-74.51, F-76.85; P-88.51, R-84.16, F-86.28; P-76.82, R-60.29, F-67.56; P-82.90, R-83.40, F-83.20
Zhang (2021)	MLT 2017, CTW1500 MSRA-TD500	GCN, RNN	P-85.19, R-90.67 F-87.85; P-83.60, R-86.45, F-85.00; P-84.54, R-86.62, F-85.57
Raisi et al. (2021)	IC15, IC17	Transformer	P-89.83, R-78.28, F-83.65; P-84.75, R-63.23, F-872.42

Table 12 The reported methods and results of the 2nd phase of deep learning-based script identification are tabulated

Researchers	Dataset	Methods	Performance
Gomez et al. (2017)	MLe2e, SIW-13, CVSI	Ensembles of conjoined frameworks	A-94.8, 94.4, 97.2
Zdenek and Nakayama (2017)	SWI-13, MLe2e, CVSI2015	CNN, SVM, MLP	A-92.83, 96.10, 97.03
Tounsi et al. (2017)	SIW-13, MLe2e, CVSI2015	Pre-trained AlexNet, VGG-16 model	A-93.9, 94.3, 98.9
Cheng et al. (2019)	SIW-13, CVSI 2015, RRC-MLT 2017	CNN, patch aggregator module	A-97.3, 98.60, 89.42
Lu et al. (2019)	SIW-13, MLe2e, CVSI-2015	ResNet-20 based framework	A-96.10, 95.80, 98.30
Bhunia et al. (2019)	SIW-13, CVSI2015, MLe2e	CNN-LSTM	A-96.50, 97.75, 96.70
Ma et al. (2021b)	RRC-MLT2017, SIW-13, CVSI-2015, MLe2e	Attention technique, Convolution based classifier	A-95.19, 96.11, 98.78, 97.20
Ghosh et al. (2021a)	CVSI-15	Lightweight CNN	A-99.06
Ghosh et al. (2022)	Bangla-384, Roman-525, Devanagari-245	CNN	A-99.30
Ghosh et al. (2021b)	Bangla-484, Roman-607, Devanagari-340	SCNN	A-99.82

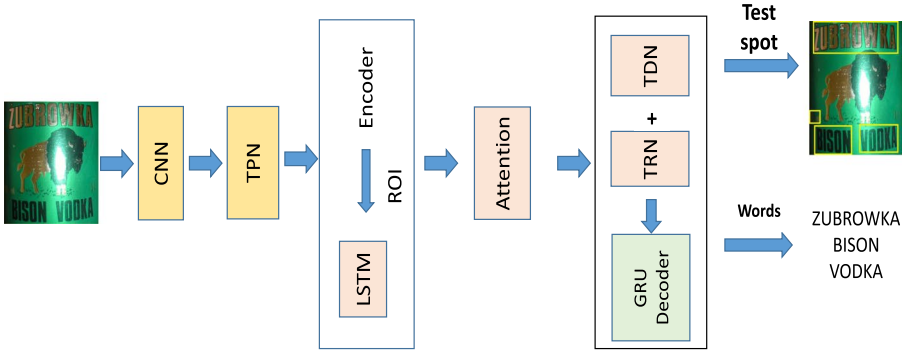

Fig. 14 End to end text detection and recognition model (Li et al. 2017)

Fig. 15 Sequence to sequence domain adaptation framework for text recognition (Zhang et al. 2019)

while working on Bangla and Devanagari text in natural images Chowdhury et al. (2011) achieved P and R of 0.72, 0.74 respectively on their developed dataset.

The works described in Sect. 4.2.2 are reported in Table 8. Apart from ICDAR 2011 dataset, testing on multi-lingual dataset Yin et al. (2013) got 63.23, 79.38, 70.39, 0.22 and 68.45, 82.63, 74.58, 0.22 of R, P, F, and speed (in s) respectively while training on 2011 and multi-lingual set respectively. The reported results of text detection using handcrafted features and machine learning classifiers of phase II are shown in Table 8.

The results of text detection work using a deep learning-based approach that is described in Sect. 4.3.1 are depicted in Table 9. Zhang et al. (2015) reported the metrics P, R, and F of 0.879, 0.908, 0.892 using 501 clusters, and considering the whole region in the image the metrics values became 0.901, 0.812, and 0.854 respectively.

The resulting summary of works in the advanced deep learning phase of Sect. 4.3.2 is presented in Table 11. Experimenting on the ICDAR2015 incident scene text dataset, Xu et al. (2019a) achieved a precision of 0.843, recall of 0.839, and F-score of 0.841. Using MSRA-TD500 they got 0.874 (P), 0.759 (R) and 0.813 (F). In 2018 Huang et al. (2019) reported that using ICDAR-2017 MLT 0.800 (P), 0.698 (R), 0.743 (F), on ICDAR-2015 0.908 (P), 0.815 (R), 0.859 (F), and using SCUT-CTW 1500 dataset 0.868 (P), 0.832 (R), 0.850 (F) results were obtained.

Table 13 Deep learning-based text recognition techniques and corresponding results as discussed in the advanced phase (2)

Researchers	Dataset	Methods	Performance
Gao et al. (2019)	IC03, IC13IIIT-5K, SVT, SVT-Perspective, CUTE80	LRA	A-98.80, 93.80, 99.60, 96.50, 79.70,75.70
Zhang et al. (2019)	IC03, IC13	SSDAN	A-92.10, 91.80
Wang et al. (2021b)	IC13, IC15Total-text, COCO Text	TRN	F-92.56, 84.23, 58.56, 33.75
Li et al. (2019a)	IC15 IC11	TPN, region feature encoder	P-80.50, R-91.40, F-85.60; P-81.70, R- 89.20, F-85.10
Naiemi et al. (2021)	IC13, IC15, IC19	CNN-based pipeline architecture	R-0.8099, P-0.9248, F-0.8635 (IC13); R-0.8309, P-0.9347, F-0.8797 (IC15); R-0.8116, P-0.9381, F-0.8703 (IC19)
Sajid et al. (2021)	IC03, IC13, IC15	Encoder centred attention framework	A-94.60, 96.30, 83.90
Raisi et al. (2020)	SVT, IC03, IC13	Transformer	A-89.34, 95.85,93.89
Atienza (2021b)	IIIT, SVT, IC03, IC13, IC15, SVTP, CT	Vision transformer	A-88.40, 87.70, 94.70, 94.30, 93.20, 92.40, 78.50, 72.60, 81.80, 81.30
Tao et al. (2021)	IIIT, SVT, IC03, IC13, IC15, SVTP, CUTe	TRIG	A-95.10, 93.8, 95.3, 95.20, 84.80, 88.10, 85.10

Table 14 Demonstration of strengths and challenges of different handcrafted and deep learning-based techniques for text detection, script identification, and text recognition

Method	Task	Strengths	Challenges
Hand-crafted based	Text detection	The MSERs technique significantly lowers the frequency of panes examined and improves the text extraction process	Connected component-based techniques are able to include a variety of non-text elements. Therefore, the effectiveness of this category of approaches depends on properly screening out the false positive rate
		The key benefit of techniques such as SWT and MSER is their computing effectiveness. This allows the detection of text elements in a single calculation with a cost of O(N), and their ability to effectively separate pixels, which makes the succeeding identification and recognition work much easier	MMS, MDS, HOG, SIFT works well in image content extraction but optimized text detection produces low performance Colour, gradient-based methods do not successfully locate the text areas
	Script identification	Feature extractors can be chosen. For time-frequency-based features, DWT, GWT, gradient, texture-related LBP, CSLBP, DLEP, etc. are used	Colour, gradient-based methods do not successfully identify all the script The key difficulty is in designing localized attributes that can manage the wide range of texts and the processing load of analyzing many blocks The handcrafted feature extractors often produce a smaller number of features thus missing some important features. The identification gives false negatives for geometrical identical characters of different categories of scripts
	Text recognition	Sliding window-based methods utilize attributes that are consolidated throughout the entire area of interest, these approaches have the advantage of being resilient to noise, distortion, and blurring	Conventional OCR methods are ineffective when employed on images of natural scenes because they are designed for the primarily gray level, stroke-based cases, as opposed to scene images where incoherent illumination, disparate fonts style and size, background complexity, noise level, and imaging obfuscations
		Dynamically picked up from bounding box identifiers without the need for specific markings Using the hybrid technique like a component-based tree-structured framework, the recognition efficiency is improved	Sliding window-based methods suffer from the generation of improper windows which results in false positives. It leads to a mismatch in the dictionary of words. Also, a high cost is required for scanning

Table 14 (continued)

Method	Task	Strengths	Challenges
Deep learning-based	Text detection	Processes any sequence, while also being detection-resistant. Be unworried about alignment, noise, and different font styles	Depend on high analytical and complicated framework layouts. inadequate recognition of overlapped or highly illuminated text
		Produce high precision and recall	For complex backgrounds, low illuminated, and curved text low performance is observed
		Generate bounding boxes on the text over compound backgrounds more quickly and effectively	
		Successfully detect multilingual, multi-oriented, and disparate-size text	
	Script identification	CNN-based methods are competent in clearly identifying the small distinctions between scripts that are challenging to differentiate, such as Chinese–Japanese, and Devanagari–Gurumukhi	To outperform other strategies, it needs a very big volume of data. Because of sophisticated data structures, training is very costlier. Deep learning also needs several workstations and pricey GPUs
		The performance of the system is high as these methods consider salient and crucial features	Since features are automatically computed, this makes it challenging to troubleshoot or comprehend why systems arrive at judgments or why specific components fail
		Attributes are dynamically determined. It is not necessary to retrieve attributes in advance	
	Text recognition	Advent of encoder-decoder, sequence to sequence, attention-based models the text recognition performance has been improved	Large data is required for training
		Can efficiently distinguish abnormal and normal textual occurrences by conducting letter categorization at every visual point	Sometimes an over-fitting problem arises
		The possibilities of absent or extraneous letters are taken into account	

Table 15 Text recognition accuracies of different typical algorithms on some benchmark datasets

Method	Year	IC03	IC13	IC15	SVT	IIITK	CUTE80	SVTP
ICDAR+ PLEX (Wang et al. 2011)	2011	57	–	–	56	–	–	–
TSM+ CRF (Shi et al. 2013)	2013	79.30	–	–	73.51	–	–	–
Whole (Goel et al. 2013)	2013	89.69	–	–	77	–	–	–
TSM+ PLEX (Shi et al. 2013)	2013	70.47			69.51	–	–	–
Label Embedding (Akata et al. 2013)	2013	–	–	–	–	76.1	–	–
PhotoOCR (Bissacco et al. 2013)	2013	–	–	–	90.39	–	–	–
Discriminative Feature Pooling (Lee et al. 2014)	2014	76	–	–	80	–	–	–
Strokelets (Yao et al. 2014c)	2014	80.33	–	–	75.89	80	–	–
Deep Features (Jaderberg et al. 2014b)	2014	91.5	–	–	86.1	–	–	–
CRNN (Shi et al. 2016b)[a]	2015	93.1	91.1	69.4	81.6	82.9	65.5	70.0
RARE (Shi et al. 2016c)[a]	2016	93.9	92.6	74.5	85.8	86.2	70.4	76.2
STAR-Net (Liu et al. 2016b)[a]	2016	94.4	92.8	76.1	86.9	87	71.7	77.5
GRCNN (Wang and Hu 2017)[a]	2017	93.5	90.9	71.4	83.7	84.2	68.1	73.6
Rosetta (Borisyuk et al. 2018)[a]	2018	93.4	90.9	71.2	84.7	84.3	69.2	73.8
Unified four stage STR (Baek et al. 2019)[a]	2019	94.9	93.6	77.6	87.5	87.9	74.0	79.2
SRN (Yu et al. 2020)	2020	-	95.5	82.7	91.5	94.8	87.8	85.1
ViTSTR- Base (Atienza 2021b)	2021	93.8	92.1	76.8	87.2	86.9	74.7	80.0
ViTSTR- Base + Aug (Atienza 2021b)	2021	94.7	93.2	78.5	87.7	88.4	81.3	81.8

[a] Trained on MJSynth (Jaderberg et al. 2014a), SynthText (Gupta et al. 2016)

4.4.3 Script identification

In 2005 Gllavata and Freisleben (2005), achieved accuracies of 83.70%, 89.00% considering Euclidean and 85.30%, 89.00% using Manhattan, and 84.50%, 89.10% by Bhattacharyya distance metrics considering K values of 5, 3, and 5 respectively in the K-NN classification scheme to identify Latin and Ideographic scripts respectively. The other results of the works as discussed in Sect. 4.2.1 are presented in Table 5.

Working on artistic scene text script identification Ghosh et al. (2019a) achieved an accuracy of 93.90% while using ELM Ghosh et al. (2020) obtained 97.95% accuracy. The results of other works as described in Sect. 4.2.2 are reported in Table 7. In 2012, Wang et al. (2012) designed a CNN framework and reported the F-score as 0.64 and 0.46 using ICDAR 2003 and SVT datasets respectively. Ma et al. (2021b) in 2021 obtained accuracies of 95.19%, 96.11%, 98.78%, and 97.20% using RRC-MLT2017, SIW-13, CVSI-2015, and MLe2e, respectively. The detailed results of deep learning based works in initial and advanced phases as described in Sects. 4.3.1 and 4.3.2 are reported in Tables 10 and 12 respectively.

4.4.4 Text recognition

Mishra et al. (2012b) achieved accuracies of 73.26%, and 81.78% using SVT, ICDAR2003. In 2013 considering the metrics such as recall, precision, and F-score, Neumann and Matas

(2013), got 45.40, 44.80, 45.20 using ICDAR 2011 dataset. The results of the research as described in Sects. 4.2.1 and 4.2.2 are tabulated in Tables 5 and 8, respectively.

Shi et al. (2016c) in 2016 obtained accuracies of 96.20%,95.50%,98.30%, 88.60% using the lexicon size of 50 experimenting with IIIT5K, SVT, ICDAR 2003, and ICDAR 2013, respectively. In 2021, Sajid et al. (2021) achieved accuracies of 94.60%, 96.30%, and 83.90% using ICDAR 2003, ICDAR 2013, and ICDAR 2015, respectively. The detailed results of deep learning-based text recognition as discussed in Sects. 4.3.1 and 4.3.2 are reported in Tables 10 and 13, respectively.

In Table 14 the merits and challenges of handcrafted and deep learning-based methods for text detection, script identification, and text recognition are presented. In Table 15 the accuracies of different methods of scene text recognition experimented on the same benchmark datasets are presented.

5 Observations

From a scientific standpoint, we conducted a foundation-based study on scene text analysis. In the last decade, several techniques were adopted and a lot of progress was seen in the field of text detection and recognition. The approaches for text localization generally can be categorized in four ways: visual/geometric/shape-based, abstract/texture-based, deep learning-based, and hybrid-based methods. The techniques like MSER, SIFT, Gabor filter, GLCM, LBP, and HOG. DCT, NMS, and SWT, etc., were the fundamental foundations of connected component analysis, sliding window-based, stroke-based, texture-based, etc., in several state-of-the-art approaches in handcrafted-based techniques. In terms of text detection and localization, the connected component approaches had a lot of achievements (Chen et al. 2011; Huang et al. 2013b; Neumann and Matas 2013; Rainarli et al. 2021; Sun et al. 2015; Yi and Tian 2011; Yin et al. 2015; Mahajan and Rani 2021). By using a quick low-level detection this process can differentiate text and non-text components. The pixels having comparable attributes were kept and subsequently combined collectively to form prospective text elements. The ERs/MSERs and the SWT were two typical approaches for this purpose also. The MSERs detector has shown that it is capable of identifying difficult text patterns and has a high recall.

To successfully narrow the search space for scene text analysis, the algorithms generally rely on connected component analysis, sliding window, texture, and stroke-based approaches. The primary focus is on retrieving the related zones in the picture text zones of interest. But this kind of approach largely depends on the identification of text-connected areas. In reality, it can be challenging to reliably identify linked text parts in scene photos with complicated backdrops, noise contamination, poor contrast, and color fluctuation. Creating a practical detector for the region of interest is likewise extremely challenging. These techniques can include a variety of non-text elements. Therefore, the effectiveness of this category of approaches depends on properly screening out the false positive rate.

The sliding window-based techniques (Wu et al. 2016) utilize a moving pane to analyze the full scene image, retrieve the potential bounding boxes, and then apply a classification algorithm to determine that the text is present inside the candidate panes. In this way, the actual text regions are found recursively. Methods relying on this approach though work well for small text areas and low-contrast images but for curved, oriented text it returns poor performance. For noisy scene text images texture-based techniques like DWT, GWT, LBP, CSLBP, GLCM, DLEP, etc., are effective.

The detection mechanism is rendered complex and ineffective by the various steps associated with conventional hand-crafted feature extraction approaches, which are also prone to erroneous sequestration. It also requires far too frequent human adjustments to categorization standards while the systems relying on deep learning retain the advantages of training algorithms. These may outperform the conventional techniques in terms of precision and effectiveness as soon as there were enough training data.

Considering the remarkable advancement of deep learning, many techniques (Nagaoka et al. 2021; Redmon and Farhadi 2017, 2018; Wang et al. 2018; Wu et al. 2019a; Yu et al. 2017; Zhao et al. 2017) were adopted to justify their effectiveness in this area. Multi-oriented, blurry, noisy, complex background text analysis in natural images has also piqued attention, owing to its greater difficulty and practicality. In terms of scene text detection, and recognition, the CNN-RNN architecture (Shi et al. 2016b), FCN (Liao et al. 2017), GAN (Kong et al. 2019), ensemble-based attention network (Gao et al. 2019), attention-based encoder-decoder (Bai et al. 2018), sequence-to-sequence attention-based network (Zhang et al. 2019), mask-R-CNN based network (Huang et al. 2019), residual attention network (Ma et al. 2021a), etc., were highly prevalent and these techniques offered significant improvements over earlier methods. In Fig. 16 the phase-wise number of papers published to the best of our knowledge for both handcrafted and deep learning-based techniques in text detection/localization, script identification, and text recognition work is depicted.

Handcrafted features with shallow learnable structures are the foundations of conventional text identification systems in natural images. By generating sophisticated combinations via obtainable various limited attributes alongside high information and machine learning-based classifiers, their efficacy quickly reaches a plateau. Such techniques often do not provide better performance (He et al. 2016b; Luo et al. 2019) compared to the deep learning methods. The universality of these low-level features on highly demanding text analysis in the wild is necessarily limited. The visual-based methods suffer from translation, scaling, or letter modification issues. In abstract techniques, the number of features or feature selection is the constraint in the performance analysis of the system.

Deep neural networks were applied in the bulk of current research items (Ganin and Lempitsky 2015; Long et al. 2015; Pei et al. 2018; Tzeng et al. 2017) to map the input and destination into a common region while the fields are synchronized. They usually try to minimize some metric of category displacement to improve the universal interpretation. Since the domain shift occurs regionally in the symbols instead of globally in the whole image as discussed in their techniques, an adaptation of such approaches conveniently to the consecutive text images where a large number of characters are involved is very difficult. To properly reallocate variable-length sequence information, character-level features are needed which were obtained by emphasizing sequence-to-sequence domain-adapted attention networks. Using the contrastive learning technique, the feature vectors were split into a series of independent pieces for self-supervised learning of sequence-to-sequence spatial identification. For low training volume data, this attention-based contrastive learning technique gives good validation accuracy.

The typical text regions in natural scene images are considered to be rectangular to help in text localization. But, instead of constructing a text mask of a predefined pattern, Mask R-CNN-driven approaches try to distinguish the text zone from the surrounding/background area. This approach estimates text boundary frames initially, then conducts lexical separation between them. In basic settings, this method is typically appropriate. Although it is somewhat unstable in the case of the predicted enclosing box which lacks encompassing the entire textual zone. Also, in the case of noisy data, this approach sometimes

Table 16 Noteworthy contributions of some state-of-the-art methods

Literature	Methodology	Noteworthy contribution	Future scope	Remarks
Yao et al. (2016)	FCN	Potential to recognize multi-oriented and bent text while also predicting the likelihood of text locations, letters, and the association between neighboring letters	The method would be able to specifically separate off letter designs from raw photos after training with higher granular attributes that could generate the succeeding text recognition process easier	Slow processing
Zhou et al. (2017)	EAST	Prediction of the correct bounding box among the closely occurring bounding boxes	Scope for improvement in performance for vertical text	FPS to be improved
Long et al. (2018)	FCN	Can detect texts of irregular forms, and text occurrences that are flat, multi-oriented, or bending	Can be extended for multi-script images	To make an adaptable approach of integrated text analysis framework
He et al. (2018)	FCN, NMS	Performs well for oriented text	Script detection for multi-script images	Improvement is needed for curved text
Lyu et al. (2018b)	ROI pooling with positional sensitivity	Using position-sensitive separation and corner point identification the localization is performed	For closely associated text segments different bounding boxes can be created	Weak performance in curved text
Bhunia et al. (2019)	Attention-based CNN-LSTM	Improved functionality for distorted, low-resolution, and complicated backgrounds	When numerous dialects share a single script, pinpointing the precise language could be an intriguing investigation topic in the future	An end-to-end method can be adopted for text detection to the identification
Cao et al. (2020)	FCOS network	Arbitrary-oriented text detection	Room for performance improvement in curved text and long text lines	Improvement in F-score
Wang et al. (2020a)	Quadrilateral region proposal network (QRPN), weighted RoI pooling, NMS	Potential to identify tiny and rotating text	Scope for bent, oriented text detection	Considers only lateral content detection
Yu et al. (2020)	SRN, GSRM	Performance is good for Lengthy non-Latin text, arbitrary text, and conventional text	For low-quality images, performance improvement is needed	Efficiency to be improved to increase its utility of it in real-world contexts

Table 16 (continued)

Literature	Methodology	Noteworthy contribution	Future scope	Remarks
Raisi et al. (2020)	Transformer	Applied transformer-based model to scene text recognition. Compared to earlier techniques, it substantially retains the spatial attributes for arbitrarily oriented text recognition	Scope for optimization for real-world applications	Inference time is higher compared to the RNN-based model
Ghosh et al. (2021b)	M-EAST, SCNN	Extracted movie titles, identified scripts, and less FPS compared to Zhou et al. (2017)	Extension of the dataset for other film industries	To consider more similarities in the texture of foreground and background
Ghosh et al. (2021a)	Lightweight CNN	Lightweight CNN was developed to deploy resource-constrained devices	The performance of the system can be improved by considering image patches for densely noisy images	For tiny and blurry images accuracy improvement is needed
Sajid et al. (2021)	Encoder centered attention framework	visual attention mechanism for parallel feature extraction with the attention process and enabled their framework to distinguish foreground and background pixels	Scope for developing newer, highly advanced fusing methods	Considerations of effectiveness and processing performance for practical implementations
Atienza (2021b)	Vision Transformer	Speedy and Effective recognition for both regular and irregular text	Scope for performance improvement in similar geometric structure in letters, vertical and curvy text	One-stage system structure which prioritizes accuracy, speed, and processing demands in proportion
Tao et al. (2021)	TRIG	Improvement in accuracy and reduce complexity	Can be extended to consider cursive text and multi-script text recognition	Inference time to process an image is 6.6 times less compared with Atienza (2021b)
Munjal et al. (2021)	TeLCoS	Text localization and script grouping at the same time to minimize the operational burden to segregate scripts	End-to-end systems can be developed for multi-script text recognition without script identification overhead	By channel pruning strategy, the system can be deployed in a low-resourced platform

Table 16 (continued)

Literature	Methodology	Noteworthy contribution	Future scope	Remarks
Wang et al. (2021b)	R-YOLO	System developed to consider oriented, slanted text detection and recognition	Scope to adapt advanced attention methods for increasing detection performance	Performance improvement is needed for tiny and curved text

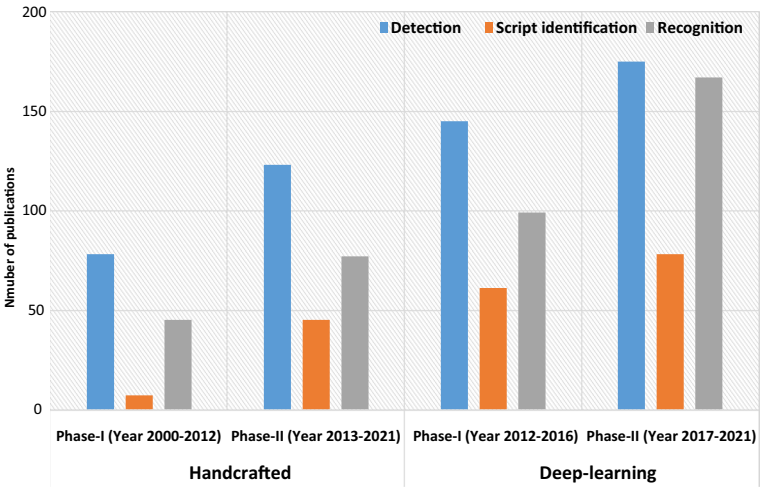

Fig. 16 Phase-wise publications based on handcrafted and deep learning techniques for scene text detection, script identification, and text recognition, respectively

generates improper text boxes. The PAN architecture considered these shortcomings of the Mask R-CNN and developed lexical separation over the word and surrounding regions by using the pyramid-level concept for the pixels in the text areas.

Region proposal structure-driven feature extraction made significant amounts of attention to developing complex detection mechanisms to get narrower bounding boxes around the text to reduce the complexity of word recognition. But, there is also the lag in vertically oriented text identification and small letter words issues in the recognition. Encode-decoder-based frameworks were developed to solve these issues. But, many studies addressed the shortcomings of the encoder-decoder platform. It was also found that there is unevenness among the ground truth symbols and the attention's output combinations (Bai et al. 2018). The probabilistic spread, which was generated by absent or redundant characters, will befuddle and misguide the training phase when contemplating the scene text recognition issue under the same attention-based encoder-decoder structure.

Observations indicate that the introduction of Transformer-based models has resulted in enhanced text detection performance in specific cases.

There are several key reasons for the performance differences between the methods based on the Transformer structure and CNN structure in the text detection tasks. Firstly, Transformers use self-attention mechanisms, which allow them to weigh the importance of different parts of the input. This is particularly useful for identifying the text regions. Secondly, Transformers can handle longer sequences of data, which is important for text detection tasks that involve processing large amounts of text. Thirdly, Transformers can incorporate the contextual information from the surrounding words to improve the accuracy of text detection. In contrast, CNNs rely on convolutional filters for identifying the text regions, and they may not capture the full context of the text. However, CNNs are useful for simpler text detection tasks, because they are more computationally efficient for smaller inputs. In Table 16 the noteworthy contributions of the state-of-the-art methods over the developing phases of scene text detection and recognition as well as the synthesis of remarkable knowledge based on the targeted literature are presented.

Presented here is an analysis of some significant works to provide insight into the trends of methods and their performances. He et al. (2018) obtained an F-score of 75.50, which is 12.29% less than the EAST model (Zhou et al. 2017) on the CASIA-10K dataset. They considered pixel-wise categorization, regression of text non-text pixels, edge coordinates, data augmentation, word, line-level annotation, etc. which according to them yielded higher scores. The F-score of PixelLink (Deng et al. 2018) is 5.5% better than EAST+PVA2x on ICDAR 2015 MLT. Using the COCO-text dataset they obtained a 2.9% higher F-score than EAST. Though PixelLink obtained a higher F-score than EAST using the same base model of VGG16, EAST provides better FPS on ICDAR 2015 MLT. Using the NVIDIA RTX 2070 of 8GB GPU, Raisi et al. (2020) got an inference time of 10 FPS and F-score of 83.65 which is higher than the method (Ma et al. 2018) for both F-score and speed. Instead of Using 1080Ti GPU which is 17% higher speed than NVIDIA RTX 2070, Wang et al. (2019) got FPS 1.6 and F-score 85.69 using the same dataset (ICDAR 2015 MLT). The hardware configurations of the state-of-the-art methods are described in the following paragraph.

5.1 Hardware configuration

The robustness (i.e., higher or lower scoring, FPS, etc.) of reported methods depends heavily upon input image resolution, system configuration, etc. Here, we discuss the hardware configuration along with the outcome of state-of-the-art techniques. Yao et al. (2016) in 2016 performed their experimentation for 480p resolution images using a K40m graphics card. The prediction time is 420ms and FPS is 1.61. The EAST (Zhou et al. 2017) method in 2017, was performed on a system with a lone NVIDIA Titan X Maxwell graphics board and an Intel E5-2670 v3 processor running at 2.30 GHz. This technique's top set has a frame rate of 16.8 FPS (considering PVANET), while its weakest configuration (considering VGG) has a frame rate of 6.52 FPS using 720p resolution. The fastest variant using PVANET2x manages only 13.2 frames per second. Ghosh et al. (2021b), proposed M-EAST which is based on the EAST technique. They experimented on a machine having a configuration of NVIDIA Quadro RTX 5000, 16 GB GPU, and primary memory (RAM) of 32GB. They reported their average FPS is 18.10 which is 4.07 times higher than EAST. Dasgupta et al. (2020) used two NVIDIAP6GPU's to train their model. The TelCos technique (Munjal et al. 2021) was run on TensorFlow 2.3 platform for 768p resolution images and the system was trained using Nvidia GeForce GTX 1080 Ti having 16GB memory. The CRAFTs (Huang et al. 2021) was trained on the Nvidia P40 GPU and Intel(R) Xeon(R) CPU. Considering 960, 1280, 1600, and 2560p resolutions the FPSs were 9.9, 8.3, 6.8, and 5.4, respectively. The FCOS-BiFPNRTX and FCOS-FPN (Cao et al. 2021) techniques were trained on TITAN GPU machine using PyTorch software. On ICDAR17 MLT the F-score of M-EAST is 84.50 while for the same dataset EAST, TelCos, FCOS-BiFPN, and FCOS-FPN, Dasgupta et al., returned 70.10, 71.13, 80.75, 78.23, 80.50, respectively. Using ICDAR 2019 MLT, M-EAST, CRAFTS, Text-spotter (Huang et al. 2021) produced F-Score of 83.08, 70.86, 72.66, respectively.

5.2 Research challenges

Text detection and recognition from natural scene photographs is a very challenging issue due to a variety of structural/topological similarities- dissimilarities, length, size of text,

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

Fig. 17 Multi-script scene text images: **a** the title of a Tollywood movie poster is written in Roman and Bangla; **b** the name of a restaurant written in Roman and Bangle at the character-level; **c** a Bollywood movie poster where the title is character-level multi-scripted in each word using Devanagari and Roman script; **d** character-level bi-script name a restaurant in Bangla and Roman script

and other factors, etc., that make this domain more interesting for researchers. Here, we discuss the challenges which need to be addressed.

- *Multi-oriented text:*Apart from horizontally aligned, the text might be diagonal, bent, circular, or even a mixture of disparate orientations to captivate the attention of onlookers. This is a major challenge for researchers to design a single technique for all types of oriented texts.
- *Background graphics:* It has been observed that the scene images are made color full and attractive. To do this the background of the image (non-text area) is designed with complicated graphics which make it difficult to extract the foreground text. Also, the foreground-background similarity makes the extraction process more challenging.
- *Issues in OCR development:* Character segregation in scene text elements is a difficult challenge owing to the vast variety of typefaces, contacting, and separated letters. As a result, despite numerous letter delineation techniques documented in the research, successful scene text element extraction, letter separation, and consequent recognition remain unsolved for making OCR engines.
- *Cam related issues:* The cameras may suffer from poor lighting circumstances, heating issues, a reflection, which makes the captured image noisy, low illuminated, etc., that complicates the scene text retrieval procedure. Also, the shaky cam often produces blurriness in images.
- *Dataset related issues:* Despite the humongous amount of datasets currently offered, contemporary handcrafted and deep learning-based strategies faces low-efficiency issue since they necessitate a sizeable amount of data with accurate ground truth for text detection and recognition.
- *Recognition efficiency:* To improve scene text recognition efficiency, attempts should be made to divide letters into single letters using disassembly algorithms to accelerate findings and assure satisfactory recognition accuracy.
- *Efficient techniques:* There is still a long way to go in terms of developing real-world methods that can accurately and reliably retrieve substantial text data in scene text. In the foreseeable future, increasingly stronger systems for text detection segmentation and recognition need to be developed to address complex datasets in actual circumstances.

- *Text length and size:* Text detection systems frequently are unable to recognize long sequences of texts, resulting in incomplete and erroneous identification. Furthermore, text occurrences containing widely separated letters can become disparate, and thus affect letter identification. Text detection in large-scale photos, on the other hand, suffers in tiny word occurrences. Furthermore, systems have been shown to adapt incoming images to different dimensions during analysis, which could significantly lower the appearance of tiny text occurrences, culminating in erroneous recognition.
- *Multilingual text:* Since multiple linguistics possess unique stroke structures (for example, Chinese and English), many present approaches for multilingual recognition and detection are ineffective. For this researchers are doing script identification first and then creating a customized framework to detect and recognize each language independently which becomes a three-stage approach (extraction, identification, recognition) and is inefficient.
- *Resource constrained environment:* With the rapid development of portable devices, real-time analysis of visual information has become extremely relevant. The main goal of this assignment is to enhance real effectiveness, precision, and resilience by optimizing running speed and saving memory, and afterward establishing an upgraded resource-constrained platform like mobile or handheld devices.
- *More complex multi-script scenarios:* Today's writing on banners and posters is multi-scripted not only on the line and word level Fig. 17a, but also on the character-level Fig. 17b–d to draw viewers attention which makes them very difficult to recognize and analyze them.

6 Future scope

The prospective approaches in text detection and localization that were found through the survey of numerous research publications are presented in this part. We have identified and assessed several issues from the previous studies that should be investigated more in the future to advance this domain.

- Several techniques asserted that they are capable of identifying and detecting curved, slanted, and perpendicular texts in scene images. But it has been observed that there are lots of failure cases in proper bounding box creation for these types of multi-oriented texts. Recently, transformer-based model (Selvam et al. 2022) has proved its efficiency in handling these issues in scene text recognition.
- The image occlusion is a major issue in localizing the texts correctly which is still lacking in the literature.
- The researchers put little concern about the situations of image capture where there are similarities in the foreground with background textures. For example, images like movie posters (Ghosh et al. 2021b) where there are analogous background textures with the text bodies are generally found. Also, artistic texts are involved in this type of poster image where creative font style, disparate font size, etc., makes hindrances in proper bounding box generation.
- The datasets need to be diverse considering uneven surfaces, dirty walls, moisture background, and low illuminated places where blurriness, noise, and low illumination are the issues that need to be addressed to make this field more realistic.

- Researchers considered both real and synthetic data in their experimentation. But, still, the synthetic data has not been considered with realistic views by varying conditions like occlusion, opacity, shadowing, dispersion, obstruction, background clutter, complex background, motion blur, etc. The advantages of synthetic data, like generalization capacity, have not yet been adequately investigated.
- In the case of text recognition, most of the work concentrated on English languages and corresponding scripts. Since the single types of scripts are handled, a script identification system wasn't necessitated in their work. However, disparate languages and scripts need to be considered to make this domain used in real-world applications. The multi-language environment is an urgent need in this domain.
- There are many geometrical similarities of letters in different scripts (e.g., Roman–Tamil, Devanagari–Gurumukhi, Devanagari–Bangla, etc.). The upcoming text recognition systems need to be efficient to address this issue.
- There are more complex real-world multi-script at character-level scene text images as presented in Fig. 17b–d that needs to be addressed.
- The major challenge faced by deep learning-based frameworks is the need for high-end processing like GPUs which restrains them to be used in low-resource platforms like mobile-based handheld devices.
- Demand for automatic scene text recognition-based system is growing in several industrial applications like tourism, the financial sector, healthcare, manufacturing, etc., as discussed in Table 1.

7 Conclusion

This study provides a detailed overview of handcrafted and deep learning-based techniques for text analysis in natural images. We began reviewing the text detection methodologies and then reported and analyzed the text categorization methods that served as the foundation for the recognition and separation systems. Following a thorough analysis of the research, it is understood how the approaches metamorphosed according to the needs/challenges/complexities in natural text images chronologically. It was observed that the efficiency of interpretation of very basic word/character images is quite good. But, as the complexity of the text strings grows due to background components, color uniformity, orientations, blurriness, noisiness, etc., the efficiency decreases. Furthermore, detection techniques are frequently not resilient to manage the related difficulties owing to camera-based distortions and integrated contextual deployments. In a summary, this article provides a wide overview of published literature to date, and also connected accessible datasets in the purview of novel ideas, standard categorization and/or recognition achievement, informative analysis, and discourse of expansive regions to investigate in sequence to accomplish the preferred objective of scene text analysis.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Aberdam A, Litman R, Tsiper S, Anschel O, Slossberg R, Mazor S, Manmatha R, Perona P (2021) Sequence-to-sequence contrastive learning for text recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 15302–15312

Afzal MZ, Pastor-Pellicer J, Shafait F, Breuel TM, Dengel A, Liwicki M (2015) Document image binarization using lstm: a sequence learning approach. In: Proceedings of the 3rd international workshop on historical document imaging and processing, pp 79–84

Agrawal P, Varma R (2012) Text extraction from images. IJCSET 2(4):1083–1087

Akata Z, Perronnin F, Harchaoui Z, Schmid C (2013) Label-embedding for attribute-based classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 819–826

Ammirato P, Berg AC (2019) A mask-rcnn baseline for probabilistic object detection. arXiv:1908.03621

Angadi S, Kodabagi M (2010) Text region extraction from low resolution natural scene images using texture features. In: 2010 IEEE 2nd international advance computing conference (IACC). IEEE, pp 121–128

Atienza R (2021a) Data augmentation for scene text recognition. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 1561–1570

Atienza R (2021b) Vision transformer for fast and efficient scene text recognition. In: International conference on document analysis and recognition. Springer, New York, pp 319–334

Azadboni MK, Samadhiya A, Khatri P (2014) Multi-orientation text detection by skeletonization (motds). In: 2014 2nd international symposium on computational and business intelligence. IEEE, pp 5–9

Baek J, Kim G, Lee J, Park S, Han D, Yun S, Oh SJ, Lee H (2019) What is wrong with scene text recognition model comparisons? Dataset and model analysis. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 4715–4723

Bai X, Shi B, Zhang C, Cai X, Qi L (2017) Text/non-text image classification in the wild with convolutional neural networks. Pattern Recogn 66:437–446

Bai F, Cheng Z, Niu Y, Pu S, Zhou S (2018) Edit probability for scene text recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1508–1516

Bhattacharyya S, Kumar J, Ghoshal K (2020) Mathematical modeling and computational tools: ICACM 2018, Kharagpur, India, November 23–25, vol 320. Springer, New York

Bhunia AK, Konwer A, Bhunia AK, Bhowmick A, Roy PP, Pal U (2019) Script identification in natural scene image and video frames using an attention based convolutional-lstm network. Pattern Recogn 85:172–184

Bissacco A, Cummins M, Netzer Y, Neven H (2013) Photoocr: reading text in uncontrolled conditions. In: Proceedings of the Ieee international conference on computer vision, pp 785–792

Borisyuk F, Gordo A, Sivakumar V (2018) Rosetta: large scale system for text detection and recognition in images. In: Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining, pp 71–79

Boureau Y-L, Bach F, LeCun Y, Ponce J (2010) Learning mid-level features for recognition. In: 2010 IEEE computer society conference on computer vision and pattern recognition. IEEE, pp 2559–2566

Busta M, Neumann L, Matas J (2017) Deep textspotter: an end-to-end trainable scene text localization and recognition framework. In: Proceedings of the IEEE international conference on computer vision, pp 2204–2212

Calvo-Zaragoza J, Gallego A-J (2019) A selectional auto-encoder approach for document image binarization. Pattern Recogn 86:37–47

Cao Y, Ma S, Pan H (2020) Fdta: fully convolutional scene text detection with text attention. IEEE Access 8:155441–155449

Cao D, Dang J, Zhong Y (2021) Towards accurate scene text detection with bidirectional feature pyramid network. Symmetry 13(3):486

Chen X, Yuille AL (2004) Detecting and reading text in natural scenes. In: Proceedings of the 2004 IEEE computer society conference on computer vision and pattern recognition. CVPR 2004, vol. 2. IEEE

Chen H, Tsai SS, Schroth G, Chen DM, Grzeszczuk R, Girod B (2011) Robust text detection in natural images with edge-enhanced maximally stable extremal regions. In: 2011 18th IEEE international conference on image processing. IEEE, pp 2609–2612

Chen J, Ma T, Xiao C (2018) Fastgcn: fast learning with graph convolutional networks via importance sampling. arXiv:1801.10247

Chen X, Jin L, Zhu Y, Luo C, Wang T (2021) Text recognition in the wild: a survey. ACM Comput Surv (CSUR) 54(2):1–35

Cheng Z, Bai F, Xu Y, Zheng G, Pu S, Zhou S (2017) Focusing attention: towards accurate text recognition in natural images. In: Proceedings of the IEEE international conference on computer vision, pp 5076–5084

Cheng Z, Xu Y, Bai F, Niu Y, Pu S, Zhou S (2018) Aon: towards arbitrarily-oriented text recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5571–5579
Cheng C, Huang Q, Bai X, Feng B, Liu W (2019) Patch aggregator for scene text script identification. In: 2019 international conference on document analysis and recognition (ICDAR). IEEE, pp 1077–1083
Ch'ng CK, Chan CS (2017) 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. IEEE, pp 935–942
Chng CK, Liu Y, Sun Y, Ng CC, Luo C, Ni Z, Fang C, Zhang S, Han J, Ding E, et al (2019) Icdar2019 robust reading challenge on arbitrary-shaped text-rrc-art. In: 2019 international conference on document analysis and recognition (ICDAR). IEEE, pp 1571–1576
Chowdhury AR, Bhattacharya U, Parui SK (2011) Text detection of two major Indian scripts in natural scene images. In: International workshop on camera-based document analysis and recognition. Springer, New York, pp 42–57
Coates A, Carpenter B, Case C, Satheesh S, Suresh B, Wang T, Wu DJ, Ng AY (2011) Text detection and character recognition in scene images with unsupervised feature learning. In: 2011 international conference on document analysis and recognition. IEEE, pp 440–445
Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05), vol 1, pp 886–893
Darab M, Rahmati M (2012) A hybrid approach to localize farsi text in natural scene images. Procedia Comput. Sci. 13:171–184
Dargan S, Kumar M, Ayyagari MR, Kumar G (2020) A survey of deep learning and its applications: a new paradigm to machine learning. Arch. Comput. Methods Eng. 27(4):1071–1092
Dasgupta K, Das S, Bhattacharya U (2020) Scale-invariant multi-oriented text detection in wild scene image. In: 2020 IEEE international conference on image processing (ICIP), pp 2041–2045. IEEE
Dauphin YN, Fan A, Auli M, Grangier D (2017) Language modeling with gated convolutional networks. In: International conference on machine learning. PMLR, pp 933–941
De Campos TE, Babu BR, Varma M et al (2009) Character recognition in natural images. VISAPP 7:1–10
Decker LGL, Pinto A, Campana JLF, Neira MC, dos Santos AA, Conceiçao JS, Angeloni MA, Li LT, et al (2020) MobText: a compact method for scene text localization. VISAPP
Del Gobbo J, Herrera RM (2020) Unconstrained text detection in manga: a new dataset and baseline. In: European conference on computer vision. Springer, New York, pp 629–646
Deng D, Liu H, Li X, Cai D (2018) Pixellink: detecting scene text via instance segmentation. In: Proceedings of the AAAI conference on artificial intelligence, vol 32
Dey S, Shivakumara P, Raghunandan K, Pal U, Lu T, Kumar GH, Chan CS (2017) Script independent approach for multi-oriented text detection in scene image. Neurocomputing 242:96–112
Dhar D, Chakraborty N, Choudhury S, Paul A, Mollah AF, Basu S, Sarkar R (2020) Multilingual scene text detection using gradient morphology. Int J Comput Vis Image Process (IJCVIP) 10(3):31–43
Dizaji KG, Zheng F, Sadoughi N, Yang Y, Deng C, Huang H (2018) Unsupervised deep generative adversarial hashing network. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3664–3673
Epshtein B, Ofek E, Wexler Y (2010) Detecting text in natural scenes with stroke width transform. In: 2010 IEEE computer society conference on computer vision and pattern recognition. IEEE, pp 2963–2970
Fang S, Xie H, Zha Z-J, Sun N, Tan J, Zhang Y (2018) Attention and language ensemble for scene text recognition with convolutional sequence modeling. In: Proceedings of the 26th ACM international conference on multimedia, pp 248–256
Fasil O, Manjunath S, Aradhya VM (2017) Word-level script identification from scene images. In: Proceedings of the 5th international conference on frontiers in intelligent computing: theory and applications. Springer, New York, pp 417–426
Feng Y, Song Y, Zhang Y (2016) Scene text detection based on multi-scale swt and edge filtering. In: 2016 23rd international conference on pattern recognition (ICPR). IEEE, pp 645–650
Fernando B, Fromont E, Tuytelaars T (2014) Mining mid-level features for image classification. Int J Comput Vision 108(3):186–203
Ganin Y, Lempitsky V (2015) Unsupervised domain adaptation by backpropagation. In: International conference on machine learning. PMLR, pp 1180–1189
Gao H, Li Y, Wang X, Han J, Li R (2019) Ensemble attention for text recognition in natural images. In: 2019 international joint conference on neural networks (IJCNN). IEEE, pp 1–8
Gao D, Li K, Wang R, Shan S, Chen X (2020) Multi-modal graph neural network for joint reasoning on vision and scene text. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 12746–12756

Garcia C, Apostolidis X (2000) Text detection and segmentation in complex color images. In: 2000 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 00CH37100), vol. 4. IEEE, pp 2326–2329
Ghosh M, Obaidullah SM, Santosh K, Das N, Roy K (2018) Artistic multi-character script identification using iterative isotropic dilation algorithm. In: International conference on recent trends in image processing and pattern recognition. Springer, New York, pp 49–62
Ghosh M, Mukherjee H, Obaidullah SM, Santosh K, Das N, Roy K (2019a) Artistic multi-character script identification. In: Document processing using machine learning. Chapman and Hall/CRC, Boston, pp 28–42
Ghosh M, Mukherjee H, Obaidullah SM, Santosh K, Das N, Roy K (2019b) Identifying the presence of graphical texts in scene images using cnn. In: 2019 international conference on document analysis and recognition workshops (ICDARW), vol 1. IEEE, pp 86–91
Ghosh M, Roy SS, Mukherjee H, Obaidullah SM, Santosh K, Roy K (2019c) Automatic text localization in scene images: a transfer learning based approach. In: National conference on computer vision, pattern recognition, image processing, and graphics. Springer, New York, pp 470–479
Ghosh M, Mukherjee H, Obaidullah SM, Santosh K, Das N, Roy K (2020) Artistic multi-script identification at character level with extreme learning machine. Procedia Comput. Sci. 167:496–505
Ghosh M, Mukherjee H, Obaidullah SM, Santosh K, Das N, Roy K (2021a) Lwsinet: a deep learning-based approach towards video script identification. Multimed Tools Appl 1:1–34
Ghosh M, Roy SS, Mukherjee H, Obaidullah SM, Gao X-Z, Roy K (2021b) Movie title extraction and script separation using shallow convolution neural network. IEEE Access 9:125184–125201
Ghosh M, Roy SS, Mukherjee H, Obaidullah SM, Santosh K, Roy K (2022) Understanding movie poster: transfer-deep learning approach for graphic-rich text recognition. Vis Comput 38(5):1645–1664
Ghoshal R, Banerjee A (2020) Svm and mlp based segmentation and recognition of text from scene images through an effective binarization scheme. In: Computational intelligence in pattern recognition. Springer, New York, pp 237–246
Girshick R (2015) Fast r-cnn. In: Proceedings of the IEEE international conference on computer vision, pp 1440–1448
Gkioxari G, Girshick R, Malik J (2015) Contextual action recognition with r* cnn. In: Proceedings of the IEEE international conference on computer vision, pp 1080–1088
Gllavata J, Ewerth R, Freisleben B (2004) Text detection in images based on unsupervised classification of high-frequency wavelet coefficients. In: Proceedings of the 17th international conference on pattern recognition, 2004. ICPR 2004, vol 1. IEEE, pp 425–428
Gllavata J, Freisleben B (2005) Script recognition in images with complex backgrounds. In: Proceedings of the fifth IEEE international symposium on signal processing and information technology, 2005. IEEE, pp 589–594
Goel V, Mishra A, Alahari K, Jawahar C (2013) Whole is greater than sum of parts: Recognizing scene text words. In: 2013 12th international conference on document analysis and recognition. IEEE, pp 398–402
Gomez L, Karatzas D (2013) Multi-script text extraction from natural scenes. In: 2013 12th international conference on document analysis and recognition. IEEE, pp 467–471
Gomez L, Karatzas D (2016) A fine-grained approach to scene text script identification. In: 2016 12th IAPR workshop on document analysis systems (DAS). IEEE, pp 192–197
Gomez L, Nicolaou A, Karatzas D (2017) Improving patch-based scene text script identification with ensembles of conjoined networks. Pattern Recogn 67:85–96
Gonzalez A, Bergasa LM, Yebes JJ, Bronte S (2012) Text location in complex images. In: Proceedings of the 21st international conference on pattern recognition (ICPR2012). IEEE, pp 617–620
Goodfellow IJ, Bulatov Y, Ibarz J, Arnoud S, Shet V (2013a) Multi-digit number recognition from street view imagery using deep convolutional neural networks. arXiv:1312.6082
Goodfellow I, Warde-Farley D, Mirza M, Courville A, Bengio Y (2013b) Maxout networks. In: International conference on machine learning. PMLR, pp 1319–1327
Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial nets. Adv Neural Inf Process Syst 27:1–10
Gupta A, Vedaldi A, Zisserman A (2016) Synthetic data for text localisation in natural images. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2315–2324
He T, Huang W, Qiao Y, Yao J (2016a) Text-attentional convolutional neural network for scene text detection. IEEE Trans Image Process 25(6):2529–2541
He K, Zhang X, Ren S, Sun J (2016b) Deep residual learning for image recognition. In: Proceedings of the ieee conference on computer vision and pattern recognition, pp 770–778

He K, Gkioxari G, Dollár P, Girshick R (2017) Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision, pp 2961–2969
He W, Zhang X-Y, Yin F, Liu C-L (2018) Multi-oriented and multi-lingual scene text detection with direct regression. IEEE Trans Image Process 27(11):5406–5419
Howe NR (2011) A Laplacian energy for document binarization. In: 2011 international conference on document analysis and recognition. IEEE, pp 6–10
Hu Z, Pi P, Wu Z, Xue Y, Shen J, Tan J, Lian X, Wang Z, Liu J (2021) E2vts: energy-efficient video text spotting from unmanned aerial vehicles. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 905–913
Huang W, Lin Z, Yang J, Wang J (2013a) Text localization in natural images using stroke feature transform and text covariance descriptors. In: Proceedings of the IEEE international conference on computer vision, pp 1241–1248
Huang R, Shivakumara P, Uchida S (2013b) Scene character detection by an edge-ray filter. In: 2013 12th international conference on document analysis and recognition. IEEE, pp 462–466
Huang W, Qiao Y, Tang X (2014) Robust scene text detection with convolution neural network induced mser trees. In: European conference on computer vision. Springer, New York, pp 497–511
Huang Z, Zhong Z, Sun L, Huo Q (2019) Mask r-cnn with pyramid attention network for scene text detection. In: 2019 IEEE winter conference on applications of computer vision (WACV). IEEE, pp 764–772
Huang J, Pang G, Kovvuri R, Toh M, Liang KJ, Krishnan P, Yin X, Hassner T (2021) A multiplexed network for end-to-end, multilingual ocr. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 4547–4557
Jaderberg M, Simonyan K, Vedaldi A, Zisserman A (2014a) Synthetic data and artificial neural networks for natural scene text recognition. arXiv:1406.2227
Jaderberg M, Vedaldi A, Zisserman A (2014b) Deep features for text spotting. In: European conference on computer vision. Springer, New York, pp 512–528
Jaderberg M, Simonyan K, Vedaldi A, Zisserman A (2016) Reading text in the wild with convolutional neural networks. Int J Comput Vis 116(1):1–20
Jang I, Ko B, Byun H, Choi Y (2002) Automatic text extraction in news images using morphology. In: Visual communications and image processing 2002, vol 4671. International Society for Optics and Photonics, pp 521–530
Juneja M, Vedaldi A, Jawahar C, Zisserman A (2013) Blocks that shout: distinctive parts for scene classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 923–930
Karatzas D, Shafait F, Uchida S, Iwamura M, Bigorda LG, Mestre SR, Mas J, Mota DF, Almazan JA, De Las Heras LP (2013) Icdar 2013 robust reading competition. In: 2013 12th international conference on document analysis and recognition. IEEE, pp 1484–1493
Karatzas D, Gomez-Bigorda L, Nicolaou A, Ghosh S, Bagdanov A, Iwamura M, Matas J, Neumann L, Chandrasekhar VR, Lu S, et al (2015) Icdar 2015 competition on robust reading. In: 2015 13th international conference on document analysis and recognition (ICDAR). IEEE, pp 1156–1160
Kasar T, Ramakrishnan AG (2011) Multi-script and multi-oriented text localization from scene images. In: International workshop on camera-based document analysis and recognition. Springer, New York, pp 1–14
Khalil A, Jarrah M, Al-Ayyoub M, Jararweh Y (2021) Text detection and script identification in natural scene images using deep learning. Comput. Electr. Eng. 91:107043
Khan T, Mollah AF (2019) Autnt-a component level dataset for text non-text classification and benchmarking with novel script invariant feature descriptors and d-cnn. Multimed Tools Appl 78(22):32159–32186
Khan T, Sarkar R, Mollah AF (2021) Deep learning approaches to scene text detection: a comprehensive review. Artif Intell Rev 54(5):3239–3298
Kim KI, Jung K, Kim JH (2003) Texture-based approach for text detection in images using support vector machines and continuously adaptive mean shift algorithm. IEEE Trans Pattern Anal Mach Intell 25(12):1631–1639
Kim K-H, Hong S, Roh B, Cheon Y, Park M (2016) Pvanet: deep but lightweight neural networks for real-time object detection. arXiv:1608.08021
Kim S, Hori T, Watanabe S (2017) Joint ctc-attention based end-to-end speech recognition using multitask learning. In: 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 4835–4839
Kipf TN, Welling M (2016) Semi-supervised classification with graph convolutional networks. arXiv:1609.02907

Kong H, Tang D, Meng X, Lu T (2019) Garn: a novel generative adversarial recognition network for end-to-end scene character recognition. In: 2019 international conference on document analysis and recognition (ICDAR). IEEE, pp 689–694
Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. Adv Neural Inf Process Syst 25:1097–1105
Kumuda T, Basavaraj L (2015) Detection and localization of text from natural scene images using texture features. In: 2015 IEEE international conference on computational intelligence and computing research (ICCIC). IEEE, pp 1–4
Lee C-Y, Bhardwaj A, Di W, Jagadeesh V, Piramuthu R (2014) Region-based discriminative feature pooling for scene text recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4050–4057
Lee CY, Baek Y, Lee H (2019) Tedeval: a fair evaluation metric for scene text detectors. In: 2019 international conference on document analysis and recognition workshops (ICDARW), vol 7. IEEE, pp 14–17
Lei Z, Zhao S, Song H, Shen J (2018) Scene text recognition using residual convolutional recurrent neural network. Mach Vis Appl 29(5):861–871
Li H, Doermann D, Kia O (2000) Automatic text detection and tracking in digital video. IEEE Trans Image Process 9(1):147–156
Li H, Wang P, Shen C (2017) Towards end-to-end text spotting with convolutional recurrent neural networks. In: Proceedings of the IEEE international conference on computer vision, pp 5238–5246
Li H, Wang P, Shen C, Zhang G (2019a) Show, attend and read: a simple and strong baseline for irregular text recognition. In: Proceedings of the AAAI conference on artificial intelligence, vol 33, pp 8610–8617
Li K, Zhang Y, Li K, Li Y, Fu Y (2019b) Visual semantic reasoning for image-text matching. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 4654–4662
Liao M, Zhang J, Wan Z, Xie F, Liang J, Lyu P, Yao C, Bai X (2019) Scene text recognition from two-dimensional perspective. In: Proceedings of the AAAI conference on artificial intelligence, vol 33, pp 8714–8721
Lim JJ, Zitnick CL, Dollár P (2013) Sketch tokens: A learned mid-level representation for contour and object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3158–3165
Lin G, Milan A, Shen C, Reid I (2017) Refinenet: multi-path refinement networks for high-resolution semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1925–1934
Lin H, Yang P, Zhang F (2020) Review of scene text detection and recognition. Arch Comput Methods Eng 27(2):433–454
Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C-Y, Berg AC (2016a) Ssd: single shot multibox detector. In: European conference on computer vision. Springer, New York, pp 21–37
Liu W, Chen C, Wong K-YK, Su Z, Han J (2016b) Star-net: a spatial attention residue network for scene text recognition. In: BMVC, vol 2, p 7
Liu Z, Lin G, Yang S, Feng J, Lin W, Goh WL (2018a) Learning markov clustering networks for scene text detection. arXiv:1805.08365
Liu Z, Li Y, Ren F, Goh WL, Yu H (2018b) Squeezedtext: a real-time scene text recognition by binary convolutional encoder-decoder network. In: Thirty-second AAAI conference on artificial intelligence
Liu X, Meng G, Pan C (2019) Scene text detection and recognition with advances in deep learning: a survey. Int J Doc Anal Recogn (IJDAR) 22(2):143–162
Liu H, Guo A, Jiang D, Hu Y, Ren B (2020) Puzzlenet: scene text detection by segment context graph learning. arXiv:2002.11371
Liu Y, He T, Chen H, Wang X, Luo C, Zhang S, Shen C, Jin L (2021) Exploring the capacity of an orderless box discretization network for multi-orientation scene text detection. Int J Comput Vis 129(6):1972–1992
Long S, He X, Yao C (2021) Scene text detection and recognition: the deep learning era. Int J Comput Vis 129(1):161–184
Long M, Cao Y, Wang J, Jordan M (2015) Learning transferable features with deep adaptation networks. In: International conference on machine learning. PMLR, pp 97–105
Long S, Ruan J, Zhang W, He X, Wu W, Yao C (2018) Textsnake: a flexible representation for detecting text of arbitrary shapes. In: Proceedings of the European conference on computer vision (ECCV), pp 20–36
Lu S, Su B, Tan CL (2010) Document image binarization using background estimation and stroke edges. Int J Doc Anal Recogn (IJDAR) 13(4):303–314

Lu L, Yi Y, Huang F, Wang K, Wang Q (2019) Integrating local CNN and global CNN for script identification in natural scene images. IEEE Access 7:52669–52679
Lucas SM (2005) Icdar 2005 text locating competition results. In: Eighth international conference on document analysis and recognition (ICDAR'05), pp 80–84
Lucas SM, Panaretos A, Sosa L, Tang A, Wong S, Young R, Ashida K, Nagai H, Okamoto M, Yamamoto H et al (2005) Icdar 2003 robust reading competitions: entries, results, and future directions. IJDAR 7(2–3):105–122
Luo C, Jin L, Sun Z (2019) Moran: a multi-object rectified attention network for scene text recognition. Pattern Recogn 90:109–118
Lyu MR, Song J, Cai M (2005) A comprehensive method for multilingual video text detection, localization, and extraction. IEEE Trans Circuits Syst Video Technol 15(2):243–255
Lyu P, Liao M, Yao C, Wu W, Bai X (2018a) 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
Lyu P, Yao C, Wu W, Yan S, Bai X (2018b) Multi-oriented scene text detection via corner localization and region segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7553–7563
Ma J, Shao W, Ye H, Wang L, Wang H, Zheng Y, Xue X (2018) Arbitrary-oriented scene text detection via rotation proposals. IEEE Trans Multimed 20(11):3111–3122
Ma C, Sun L, Zhong Z, Huo Q (2021a) Relatext: exploiting visual relationships for arbitrary-shaped scene text detection with graph convolutional networks. Pattern Recogn 111:107684
Ma M, Wang Q-F, Huang S, Huang S, Goulermas Y, Huang K (2021b) Residual attention-based multi-scale script identification in scene text images. Neurocomputing 421:222–233
Mafla A, Dey S, Biten AF, Gomez L, Karatzas D (2021) Multi-modal reasoning graph for scene-text based fine-grained image classification and retrieval. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp 4023–4033
Mahajan S, Rani R (2021) Text detection and localization in scene images: a broad review. Artif Intell Rev 54(6):4317–4377
Mathew M, Jain M, Jawahar C (2017) Benchmarking scene text recognition in devanagari, telugu and malayalam. In: 2017 14th IAPR international conference on document analysis and recognition (ICDAR), vol 7. IEEE, pp 42–46
Mei J, Dai L, Shi B, Bai X (2016) Scene text script identification with convolutional recurrent neural networks. In: 2016 23rd international conference on pattern recognition (ICPR). IEEE, pp 4053–4058
Mishra A, Alahari K, Jawahar C (2012a) Scene text recognition using higher order language priors. In: BMVC-British Machine Vision Conference. BMVA
Mishra A, Alahari K, Jawahar C (2012b) Top-down and bottom-up cues for scene text recognition. In: 2012 IEEE conference on computer vision and pattern recognition, pp 2687–2694
Munjal RS, Goyal M, Moharir R, Moharana S (2021) Telcos: on device text localization with clustering of script. arXiv:2104.08045
Nagaoka Y, Miyazaki T, Sugaya Y, Omachi S (2021) Text detection using multi-stage region proposal network sensitive to text scale. Sensors 21(4):1232
Naiemi F, Ghods V, Khalesi H (2021) A novel pipeline framework for multi oriented scene text image detection and recognition. Expert Syst Appl 170:114549
Nayef N, Yin F, Bizid I, Choi H, Feng Y, Karatzas D, Luo Z, Pal U, Rigaud C, Chazalon J, et al (2017) Icdar2017 robust reading challenge on multi-lingual scene text detection and script identification-rrc-mlt. In: 2017 14th IAPR international conference on document analysis and recognition (ICDAR), vol 1. IEEE, pp 1454–1459
Nayef N, Patel Y, Busta M, Chowdhury PN, Karatzas D, Khlif W, Matas J, Pal U, Burie J-C, Liu C-l, et al (2019) Icdar2019 robust reading challenge on multi-lingual scene text detection and recognition-rrc-mlt-2019. In: 2019 international conference on document analysis and recognition (ICDAR). IEEE, pp 1582–1587
Neumann L, Matas J (2010) A method for text localization and recognition in real-world images. In: Asian conference on computer vision. Springer, New York, pp 770–783
Neumann L, Matas J (2012) Real-time scene text localization and recognition. In: 2012 ieee conference on computer vision and pattern recognition. IEEE, pp 3538–3545
Neumann L, Matas J (2013) Scene text localization and recognition with oriented stroke detection. In: Proceedings of the Ieee international conference on computer vision, pp 97–104
Otsu N (1979) A threshold selection method from gray-level histograms. IEEE Trans Syst Man Cybern 9(1):62–66

Pan Y-F, Hou X, Liu C-L (2009) Text localization in natural scene images based on conditional random field. In: 2009 10th international conference on document analysis and recognition. IEEE, pp 6–10
Pan Y-F, Liu C-L, Hou X (2010a) Fast scene text localization by learning-based filtering and verification. In: 2010 IEEE international conference on image processing. IEEE, pp 2269–2272
Pan Y-F, Hou X, Liu C-L (2010b) A hybrid approach to detect and localize texts in natural scene images. IEEE Trans Image Process 20(3):800–813
Pandey D, Pandey BK, Wairya S (2021) Hybrid deep neural network with adaptive galactic swarm optimization for text extraction from scene images. Soft Comput 25(2):1563–1580
Pastor-Pellicer J, España-Boquera S, Zamora-Martínez F, Afzal MZ, Castro-Bleda MJ (2015) Insights on the use of convolutional neural networks for document image binarization. In: International work-conference on artificial neural networks. Springer, New York, pp 115–126
Paul S, Saha S, Basu S, Saha PK, Nasipuri M (2019) Text localization in camera captured images using fuzzy distance transform based adaptive stroke filter. Multimed Tools Appl 78(13):18017–18036
Pei Z, Cao Z, Long M, Wang J (2018) Multi-adversarial domain adaptation. In: Thirty-second AAAI conference on artificial intelligence
Peng X, Cao H, Natarajan P (2017) Using convolutional encoder-decoder for document image binarization. In: 2017 14th IAPR international conference on document analysis and recognition (ICDAR), vol 1. IEEE, pp 708–713
Phan TQ, Shivakumara P, Ding Z, Lu S, Tan CL (2011) Video script identification based on text lines. In: 2011 international conference on document analysis and recognition. IEEE, pp 1240–1244
Phan TQ, Shivakumara P, Tan CL (2012) Detecting text in the real world. In: Proceedings of the 20th ACM international conference on multimedia, pp 765–768
Phan TQ, Shivakumara P, Tian S, Tan CL (2013) Recognizing text with perspective distortion in natural scenes. In: Proceedings of the IEEE international conference on computer vision, pp 569–576
Pratikakis I, Gatos B, Ntirogiannis K (2013) Icdar 2013 document image binarization contest (dibco 2013). In: 2013 12th international conference on document analysis and recognition. IEEE, pp 1471–1476
Qin X, Jiang J, Yuan C-A, Qiao S, Fan W (2020) Arbitrary shape natural scene text detection method based on soft attention mechanism and dilated convolution. IEEE Access 8:122685–122694
Raghunandan K, Shivakumara P, Roy S, Kumar GH, Pal U, Lu T (2018) Multi-script-oriented text detection and recognition in video/scene/born digital images. IEEE Trans Circuits Syst Video Technol 29(4):1145–1162
Rainarli E et al (2021) A decade: review of scene text detection methods. Comput. Sci. Rev. 42:100434
Raisi Z, Naiel MA, Fieguth P, Wardell S, Zelek J (2020) 2d positional embedding-based transformer for scene text recognition. J Comput Vis Imaging Syst 6(1):1–4
Raisi Z, Naiel MA, Younes G, Wardell S, Zelek JS (2021) Transformer-based text detection in the wild. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 3162–3171
Rashmi V, Nayak SN (2018) A hybrid approach to localize text in natural scene images. Int J Eng Appl Sci Technol 3(1):53–60
Redmon J, Farhadi A (2017) Yolo9000: better, faster, stronger. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7263–7271
Redmon J, Farhadi A (2018) Yolov3: an incremental improvement. arXiv:1804.02767
Ren X, Ramanan D (2013) Histograms of sparse codes for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3246–3253
Risnumawan A, Shivakumara P, Chan CS, Tan CL (2014) A robust arbitrary text detection system for natural scene images. Expert Syst Appl 41(18):8027–8048
Risnumawan A, Sulistijono IA, Abawajy J (2016) Text detection in low resolution scene images using convolutional neural network. In: International conference on soft computing and data mining. Springer, New York, pp 366–375
Sajid U, Chow M, Zhang J, Kim T, Wang G (2021) Parallel scale-wise attention network for effective scene text recognition. arXiv:2104.12076
Selvam P, Koilraj JAS, Romero CAT, Alharbi M, Mehbodniya A, Webber JL, Sengan S (2022) A transformer-based framework for scene text recognition. IEEE Access 10:100895–100910
Sengupta P, Mollah AF (2021) Scene character recognition with morphological filtering and hog features. In: Soft computing techniques and applications. Springer, New York, pp 1–9
Sermanet P, Chintala S, LeCun Y (2012) Convolutional neural networks applied to house numbers digit classification. In: Proceedings of the 21st international conference on pattern recognition (ICPR2012). IEEE, pp 3288–3291
Shahab A, Shafait F, Dengel A (2011) Icdar 2011 robust reading competition challenge 2: Reading text in scene images. In: 2011 international conference on document analysis and recognition, pp 1491–1496

Sharma N, Mandal R, Sharma R, Pal U, Blumenstein M (2015) Icdar2015 competition on video script identification (cvsi 2015). In: 2015 13th international conference on document analysis and recognition (ICDAR). IEEE, pp 1196–1200
Sheng F, Chen Z, Xu B (2019) Nrtr: a no-recurrence sequence-to-sequence model for scene text recognition. In: 2019 international conference on document analysis and recognition (ICDAR). IEEE, pp 781–786
Shi C, Xiao B, Wang C, Zhang Y (2012) Graph-based background suppression for scene text detection. In: 2012 10th IAPR international workshop on document analysis systems. IEEE, pp 210–214
Shi C, Wang C, Xiao B, Zhang Y, Gao S, Zhang Z (2013) Scene text recognition using part-based tree-structured character detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2961–2968
Shi B, Yao C, Zhang C, Guo X, Huang F, Bai X (2015) Automatic script identification in the wild. In: 2015 13th international conference on document analysis and recognition (ICDAR). IEEE, pp 531–535
Shi B, Bai X, Yao C (2016a) Script identification in the wild via discriminative convolutional neural network. Pattern Recogn 52:448–458
Shi B, Bai X, Yao C (2016b) 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
Shi B, Wang X, Lyu P, Yao C, Bai X (2016c) Robust scene text recognition with automatic rectification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4168–4176
Shi B, Bai X, Belongie S (2017a) Detecting oriented text in natural images by linking segments. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2550–2558
Shi B, Yao C, Liao M, Yang M, Xu P, Cui L, Belongie S, Lu S, Bai X (2017b) Icdar2017 competition on reading Chinese text in the wild (rctw-17). In: 2017 14th IAPR international conference on document analysis and recognition (ICDAR), vol 1. IEEE, pp 1429–1434
Shi B, Yang M, Wang X, Lyu P, Yao C, Bai X (2018) Aster: an attentional scene text recognizer with flexible rectification. IEEE Trans Pattern Anal Mach Intell 41(9):2035–2048
Shinde A, Patil M (2021) Street view text detection methods. In: 2021 international conference on artificial intelligence and smart systems (ICAIS). IEEE, pp 961–965
Shivakumara P, Phan TQ, Tan CL (2010) A Laplacian approach to multi-oriented text detection in video. IEEE Trans Pattern Anal Mach Intell 33(2):412–419
Shivakumara P, Sreedhar RP, Phan TQ, Lu S, Tan CL (2012) Multioriented video scene text detection through Bayesian classification and boundary growing. IEEE Trans Circuits Syst Video Technol 22(8):1227–1235
Shivakumara P, Yuan Z, Zhao D, Lu T, Tan CL (2015) New gradient-spatial-structural features for video script identification. Comput Vis Image Underst 130:35–53
Simanjuntak GD, Nugroho H (2021) Scene text detection with quadtree-based candidate text regions and convolutional neural network. Int J Electr Eng Inf 13(1):152–162
Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556
Singh AK, Mishra A, Dabral P, Jawahar C (2016) A simple and effective solution for script identification in the wild. In: 2016 12th IAPR workshop on document analysis systems (DAS). IEEE, pp 428–433
Soni R, Kumar B, Chand S (2019) Text detection and localization in natural scene images based on text awareness score. Appl Intell 49(4):1376–1405
Sravani M, Maheswararao A, Murthy MK (2021) Robust detection of video text using an efficient hybrid method via key frame extraction and text localization. Multimed Tools Appl 80(6):9671–9686
Sriman B, Schomaker L (2019) Multi-script text versus non-text classification of regions in scene images. J Vis Commun Image Represent 62:23–42
Su B, Lu S (2014) Accurate scene text recognition based on recurrent neural network. In: Asian conference on computer vision. Springer, New York, pp 35–48
Su Y-M, Peng H-W, Huang K-W, Yang C-S (2019) Image processing technology for text recognition. In: 2019 international conference on technologies and applications of artificial intelligence (TAAI). IEEE, pp 1–5
Sun L, Huo Q, Jia W, Chen K (2015) A robust approach for text detection from natural scene images. Pattern Recogn 48(9):2906–2920
Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1–9
Tang J, Yang Z, Wang Y, Zheng Q, Xu Y, Bai X (2019) Seglink++: detecting dense and arbitrary-shaped scene text by instance-aware component grouping. Pattern Recogn 96:106954
Tao Y, Jia Z, Ma R, Xu S (2021) Trig: transformer-based text recognizer with initial embedding guidance. Electronics 10(22):2780

Tian Z, Shen C, Chen H, He T (2019) Fcos: fully convolutional one-stage object detection. In: Proceedings of the IEEE/cvf international conference on computer vision, pp 9627–9636
Tounsi M, Moalla I, Lebourgeois F, Alimi AM (2017) Cnn based transfer learning for scene script identification. In: International conference on neural information processing. Springer, New York, pp 702–711
Turki H, Halima MB, Alimi AM (2016) Text detection in natural scene images using two masks filtering. In: 2016 IEEE/ACS 13th international conference of computer systems and applications (AICCSA). IEEE, pp 1–6
Turki H, Halima MB, Alimi AM (2017) A hybrid method of natural scene text detection using msers masks in hsv space color. In: Ninth international conference on machine vision (ICMV 2016), vol 10341. International Society for Optics and Photonics, p 1034111
Tzeng E, Hoffman J, Saenko K, Darrell T (2017) Adversarial discriminative domain adaptation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7167–7176
Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. Adv Neural Inf Process Syst 30:1–10
Veit A, Matera T, Neumann L, Matas J, Belongie S (2016) Coco-text: dataset and benchmark for text detection and recognition in natural images. arXiv:1601.07140
Verma M, Sood N, Roy PP, Raman B (2017) Script identification in natural scene images: a dataset and texture-feature based performance evaluation. In: Proceedings of international conference on computer vision and image processing. Springer, New York, pp 309–319
Wang K, Belongie S (2010) Word spotting in the wild. In: European conference on computer vision. Springer, New York, pp 591–604
Wang J, Hu X (2017) Gated recurrent convolution neural network for ocr. Adv Neural Inf Process Syst 30:1–10
Wang K, Babenko B, Belongie S (2011) End-to-end scene text recognition. In: 2011 international conference on computer vision. IEEE, pp 1457–1464
Wang T, Wu DJ, Coates A, Ng AY (2012) End-to-end text recognition with convolutional neural networks. In: Proceedings of the 21st international conference on pattern recognition (ICPR2012). IEEE, pp 3304–3308
Wang X, Wang B, Bai X, Liu W, Tu Z (2013) Max-margin multiple-instance dictionary learning. In: International conference on machine learning, pp 846–854
Wang P, Chen P, Yuan Y, Liu D, Huang Z, Hou X, Cottrell G (2018) Understanding convolution for semantic segmentation. In: 2018 IEEE winter conference on applications of computer vision (WACV), pp 1451–1460. IEEE
Wang W, Xie E, Li X, Hou W, Lu T, Yu G, Shao S (2019) Shape robust text detection with progressive scale expansion network. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 9336–9345
Wang S, Liu Y, He Z, Wang Y, Tang Z (2020a) A quadrilateral scene text detector with two-stage network architecture. Pattern Recogn 102:107230
Wang T, Zhu Y, Jin L, Luo C, Chen X, Wu Y, Wang Q, Cai M (2020b) Decoupled attention network for text recognition. In: Proceedings of the AAAI conference on artificial intelligence, vol 34, pp 12216–12224
Wang X, Zheng S, Zhang C, Li R, Gui L (2021a) R-yolo: a real-time text detector for natural scenes with arbitrary rotation. Sensors 21(3):888
Wang P, Li H, Shen C (2021b) Towards end-to-end text spotting in natural scenes. IEEE Trans Pattern Anal Mach Intell
Wojna Z, Gorban AN, Lee D-S, Murphy K, Yu Q, Li Y, Ibarz J (2017) Attention-based extraction of structured information from street view imagery. In: 2017 14th IAPR international conference on document analysis and recognition (ICDAR), vol 1. IEEE, pp 844–850
Wolf C, Doermann D (2002) Binarization of low quality text using a markov random field model. In: Object recognition supported by user interaction for service robots, vol 3. IEEE, pp 160–163
Wolf C, Jolion J-M (2006) Object count/area graphs for the evaluation of object detection and segmentation algorithms. IJDAR 8(4):280–296
Wu H, Zou B, Zhao Y-Q, Chen Z, Zhu C, Guo J (2016) Natural scene text detection by multi-scale adaptive color clustering and non-text filtering. Neurocomputing 214:1011–1025
Wu F, Souza A, Zhang T, Fifty C, Yu T, Weinberger K (2019a) Simplifying graph convolutional networks. In: International conference on machine learning. PMLR, pp 6861–6871
Wu H, Zhang J, Huang K, Liang K, Yu Y (2019b) Fastfcn: rethinking dilated convolution in the backbone for semantic segmentation. arXiv:1903.11816
Xie S, Tu Z (2015) Holistically-nested edge detection. In: Proceedings of the IEEE international conference on computer vision, pp 1395–1403
Xu Y, Wang Y, Zhou W, Wang Y, Yang Z, Bai X (2019a) Textfield: learning a deep direction field for irregular scene text detection. IEEE Trans Image Process 28(11):5566–5579

Xu H, Su X, Liu T, Guo P, Gao G, Bao F (2019b) A natural scene text extraction approach based on generative adversarial learning. In: International conference on neural information processing. Springer, New York, pp 65–73
Yang J, Yu K, Gong Y, Huang T (2009) Linear spatial pyramid matching using sparse coding for image classification. In: 2009 IEEE conference on computer vision and pattern recognition. IEEE, pp 1794–1801
Yang X, He D, Zhou Z, Kifer D, Giles CL (2017) Learning to read irregular text with attention mechanisms. In: IJCAI, vol 1, p 3
Yang B, Ma AJ, Yuen PC (2018) Learning domain-shared group-sparse representation for unsupervised domain adaptation. Pattern Recogn 81:615–632
Yang M, Guan Y, Liao M, He X, Bian K, Bai S, Yao C, Bai X (2019) Symmetry-constrained rectification network for scene text recognition. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 9147–9156
Yao C, Bai X, Liu W, Ma Y, Tu Z (2012) Detecting texts of arbitrary orientations in natural images. In: 2012 IEEE conference on computer vision and pattern recognition. IEEE, pp 1083–1090
Yao C, Bai X, Liu W (2014a) A unified framework for multioriented text detection and recognition. IEEE Trans Image Process 23(11):4737–4749
Yao C, Bai X, Shi B, Liu W (2014b) Strokelets: a learned multi-scale representation for scene text recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4042–4049
Yao C, Bai X, Shi B, Liu W (2014c) Strokelets: a learned multi-scale representation for scene text recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4042–4049
Yao C, Wu J, Zhou X, Zhang C, Zhou S, Cao Z, Yin Q (2015) Incidental scene text understanding: Recent progresses on icdar 2015 robust reading competition challenge 4. arXiv:1511.09207
Yao C, Bai X, Sang N, Zhou X, Zhou S, Cao Z (2016) Scene text detection via holistic, multi-channel prediction. arXiv:1606.09002
Yi C, Tian Y (2011) Text string detection from natural scenes by structure-based partition and grouping. IEEE Trans Image Process 20(9):2594–2605
Yi C, Tian Y (2013) Text extraction from scene images by character appearance and structure modeling. Comput Vis Image Underst 117(2):182–194
Yildirim G, Achanta R, Süsstrunk S (2013) Text recognition in natural images using multiclass hough forests. In: Proceedings of the 8th international conference on computer vision theory and applications, vol 1, pp 737–741
Yin X, Yin X-C, Hao H-W, Iqbal K (2012) Effective text localization in natural scene images with mser, geometry-based grouping and adaboost. In: Proceedings of the 21st international conference on pattern recognition (ICPR2012). IEEE, pp 725–728
Yin X-C, Yin X, Huang K, Hao H-W (2013) Robust text detection in natural scene images. IEEE Trans Pattern Anal Mach Intell 36(5):970–983
Yin X-C, Pei W-Y, Zhang J, Hao H-W (2015) Multi-orientation scene text detection with adaptive clustering. IEEE Trans Pattern Anal Mach Intell 37(9):1930–1937
Yu F, Koltun V, Funkhouser T (2017) Dilated residual networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 472–480
Yu D, Li X, Zhang C, Liu T, Han J, Liu J, Ding E (2020) Towards accurate scene text recognition with semantic reasoning networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 12113–12122
Yuliang L, Lianwen J, Shuaitao Z, Sheng Z (2017) Detecting curve text in the wild: new dataset and new solution. arXiv:1712.02170
Zdenek J, Nakayama H (2017) Bag of local convolutional triplets for script identification in scene text. In: 2017 14th IAPR international conference on document analysis and recognition (ICDAR), vol 1. IEEE, pp 369–375
Zeiler MD (2012) Adadelta: an adaptive learning rate method. arXiv:1212.5701
Zhan F, Lu S (2019) Esir: end-to-end scene text recognition via iterative image rectification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 2059–2068
Zhang C, Yao C, Shi B, Bai X (2015) Automatic discrimination of text and non-text natural images. In: 2015 13th International conference on document analysis and recognition (icdar). IEEE, pp 886–890
Zhang S, Liu Y, Jin L, Luo C (2018) Feature enhancement network: a refined scene text detector. In: Proceedings of the AAAI conference on artificial intelligence, vol 32
Zhang Y, Nie S, Liu W, Xu X, Zhang D, Shen HT (2019) Sequence-to-sequence domain adaptation network for robust text image recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 2740–2749

Zhang S-X, Zhu X, Yang C, Wang H, Yin X-C (2021a) Adaptive boundary proposal network for arbitrary shape text detection. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 1305–1314

Zhang M, Ma M, Wang P (2021b) Scene text recognition with cascade attention network. In: Proceedings of the 2021 international conference on multimedia retrieval, pp 385–393

Zhao D, Shivakumara P, Lu S, Tan CL (2012) New spatial-gradient-features for video script identification. In: 2012 10th IAPR international workshop on document analysis systems. IEEE, pp 38–42

Zhao H, Shi J, Qi X, Wang X, Jia J (2017) Pyramid scene parsing network. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2881–2890

Zharikov I, Nikitin P, Vasiliev I, Dokholyan V (2020) Ddi-100: Dataset for text detection and recognition. In: Proceedings of the 2020 4th international symposium on computer science and intelligent control, pp 1–5

Zhou X, Yao C, Wen H, Wang Y, Zhou S, He W, Liang J (2017a) East: an efficient and accurate scene text detector. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5551–5560

Zhuo J, Wang S, Zhang W, Huang Q (2017b) Deep unsupervised convolutional domain adaptation. In: Proceedings of the 25th ACM international conference on multimedia, pp 261–269

Zhu Y, Du J (2021) Textmountain: accurate scene text detection via instance segmentation. Pattern Recogn 110:107336

Zhu X, Zhang Z (2021) Transformer-based end-to-end scene text recognition. In: 2021 IEEE 16th conference on industrial electronics and applications (ICIEA), pp 1691–1695. https://doi.org/10.1109/ICIEA51954.2021.9516154

Terms and Conditions

Springer Nature journal content, brought to you courtesy of Springer Nature Customer Service Center GmbH ("Springer Nature").

Springer Nature supports a reasonable amount of sharing of research papers by authors, subscribers and authorised users ("Users"), for small-scale personal, non-commercial use provided that all copyright, trade and service marks and other proprietary notices are maintained. By accessing, sharing, receiving or otherwise using the Springer Nature journal content you agree to these terms of use ("Terms"). For these purposes, Springer Nature considers academic use (by researchers and students) to be non-commercial.

These Terms are supplementary and will apply in addition to any applicable website terms and conditions, a relevant site licence or a personal subscription. These Terms will prevail over any conflict or ambiguity with regards to the relevant terms, a site licence or a personal subscription (to the extent of the conflict or ambiguity only). For Creative Commons-licensed articles, the terms of the Creative Commons license used will apply.

We collect and use personal data to provide access to the Springer Nature journal content. We may also use these personal data internally within ResearchGate and Springer Nature and as agreed share it, in an anonymised way, for purposes of tracking, analysis and reporting. We will not otherwise disclose your personal data outside the ResearchGate or the Springer Nature group of companies unless we have your permission as detailed in the Privacy Policy.

While Users may use the Springer Nature journal content for small scale, personal non-commercial use, it is important to note that Users may not:

1. use such content for the purpose of providing other users with access on a regular or large scale basis or as a means to circumvent access control;
2. use such content where to do so would be considered a criminal or statutory offence in any jurisdiction, or gives rise to civil liability, or is otherwise unlawful;
3. falsely or misleadingly imply or suggest endorsement, approval , sponsorship, or association unless explicitly agreed to by Springer Nature in writing;
4. use bots or other automated methods to access the content or redirect messages
5. override any security feature or exclusionary protocol; or
6. share the content in order to create substitute for Springer Nature products or services or a systematic database of Springer Nature journal content.

In line with the restriction against commercial use, Springer Nature does not permit the creation of a product or service that creates revenue, royalties, rent or income from our content or its inclusion as part of a paid for service or for other commercial gain. Springer Nature journal content cannot be used for inter-library loans and librarians may not upload Springer Nature journal content on a large scale into their, or any other, institutional repository.

These terms of use are reviewed regularly and may be amended at any time. Springer Nature is not obligated to publish any information or content on this website and may remove it or features or functionality at our sole discretion, at any time with or without notice. Springer Nature may revoke this licence to you at any time and remove access to any copies of the Springer Nature journal content which have been saved.

To the fullest extent permitted by law, Springer Nature makes no warranties, representations or guarantees to Users, either express or implied with respect to the Springer nature journal content and all parties disclaim and waive any implied warranties or warranties imposed by law, including merchantability or fitness for any particular purpose.

Please note that these rights do not automatically extend to content, data or other material published by Springer Nature that may be licensed from third parties.

If you would like to use or distribute our Springer Nature journal content to a wider audience or on a regular basis or in any other manner not expressly permitted by these Terms, please contact Springer Nature at

onlineservice@springernature.com