

International Journal of Knowledge Processing Studies (KPS)



Homepage: <http://kps.artahub.ir/>



ORIGINAL RESEARCH ARTICLE

Deep Learning Approaches for Multilingual Text Detection and Script Recognition in Video Scenes: A Comprehensive Survey

Rituraj Jain¹, Kamal Upreti^{2*}, Rajesh Kumar³, Dyuti Banerjee⁴, Shitiz Upreti⁵, Ramesh Babu P⁶

¹ School of Computer Science Engineering, Chandigarh University Uttar Pradesh, Unnao, U.P., India, jainrituraj@yahoo.com

² Department of Computer Science, CHRIST University, Delhi-NCR Ghaziabad, Uttar Pradesh, India, kamalupreti1989@gmail.com, 0000-0003-0665-530X.

³ Department of Computer Science Engineering -AI-ML, Meerut institute of Engineering and technology, Meerut, India, rajesh.kumar@miet.ac.in

⁴ Artificial Intelligence and Data Science Department, Koneru Lakshmaiah Education Foundation, Andhra Pradesh, India, dyuti738900@gmail.com

⁵ Centre of Distance and Online Education, Manav Rachna International Institute of Research & Studies, Faridabad, India, upretiec@gmail.com

⁶ Department of CSE, Department of Computer Science & Engineering, Narsimha Reddy Engineering College, Secunderabad, Telangana, India, drprb2009@gmail.com

ARTICLE INFO

Article History:

Received: 2026-05-14

Revised: 2026-06-22

Accepted: 2026-07-22

Published Online: 2026-09-01

Keywords:

Computer Vision, Convolutional Neural Networks, Recurrent Neural Networks, Transformers, Optical Character Recognition.

Number of Reference: 94

Number of Figures: 7

Number of Tables: 11

DOI: 10.22034/kps.2026.595307.1294



ABSTRACT

The growing demand for multilingual text and script detection in video scenes across media, education, and surveillance requires accurate real-time recognition despite challenges such as dynamic environments, diverse scripts, low-resource languages, complex layouts, and ethical concerns. This review compares CNNs, RNNs, transformers, and hybrid models in terms of performance, advantages, limitations, and applications. A systematic review of studies published from 2012 to 2024 was conducted using IEEE Xplore, ACM Digital Library, and Google Scholar. Studies were included based on relevance to deep learning-based multilingual text detection, text recognition, or script recognition in video-scene environments and the availability of quantitative evaluation results. Findings highlight the potential of hybrid architectures, particularly Mask TextSpotter v3, for end-to-end text spotting. However, cross-study comparison is limited by variations in datasets, evaluation protocols, and reporting practices. A total of 58 studies met the eligibility criteria. Key trends include cross-lingual transfer learning, multimodal integration, and contextual understanding, while future directions include self-supervised learning, privacy-aware recognition, and NLP-integrated systems. ©authors.

1. Introduction

Video scene analysis as a technology has therefore become invaluable especially in today's increasingly visual environment. Be it in strengthening the safety of self-driving cars or more effective content filtering and more readily available subtitles for video content, the extraction of meaningful information from videos comes into highest demand (Dong et al., 2022). Since the core technologies of this field such as computer vision, machine learning, and artificial intelligence have emerged in recent years, this field has expanded dramatically. As extended uses, applications are found in surveillance and security, analysis of sports performance, and in providing richer media experiences (Moeslund et al., 2017).

In this review, different text processing tasks are distinguished as follows. Text detection refers to identifying and localizing text regions within images or video frames. Text recognition refers to converting detected text regions into readable character sequences. Script identification refers to determining the writing system of the detected text, such as Latin, Arabic, Chinese, or Devanagari, whereas script recognition refers to recognizing textual content across different writing systems. Language identification refers to determining the language associated with the detected or recognized text. Although these tasks are closely related and may be integrated within end-to-end systems, they represent distinct processing stages and should not be used interchangeably.

Within this broader field, the detection and recognition of text in multilingual and multi-script environment are a challenging issue. Issues such as font differentiations based on language, script, and writing style or right to left text orientation, background complexity, and camera motion make this task complex. Real-time processing conditions pose even higher risks to the needless misinterpretation of meaning and its subsequent detrimental consequences for system performance, especially when processing varies contexts of language and vision (Merino- Gracia & Mirmehdi, 2014; Sassu et al., 2021).

Deep learning has recently transformed video scene analysis, specifically, multilingual text detection and script recognition. Since deep neural networks are capable of learning hierarchical feature representations, DNNs have shown great enhancements over conventional approaches that barely enabled addressing the feature complexity of video environments (X. Liu et al., 2019). Recent architectures like Convolutional Neural Network (CNNs), Recurrent Neural Network (RNNs) and transformers have contributed a lot to text detection and recognition methods (Mehrish et al., 2023; Mirza et al., 2020).

With the growing number of videos being created in recent years, there are explosive requirements on video scenes related to multilingual text detection and script recognition in various applications such as media, educational, and surveillance. However, challenges that remain, such as dynamic environments and diversity of scripts, limited languages and complexity of the layout for challenges related to real-time processing, scalability, and robustness. In many such instances, traditional approaches simply don't work—and in particular for occlusion, motion blur or multiple-language text scenarios. As such, the current limitations are being addressed using deep learning techniques, and a much more efficient, real-time scalable solution to this problem, namely video scene analysis on multilingual scenes, is achieved.

In this paper, we focus on the recent developments in multilingual text detection and recognition in video scene scripts using deep learning. State-of-the-art architectures, such as CNNs and RNNs, transformers and hybrid models together with their applications, advantages and limitations will also be covered. It tackles some of the basic challenges in the area, including low resolution, motion blur, cross-lingual adaptation, and highlights new trends including cross-lingual transfer learning, multimodal data integration and self-supervised learning. The aim of this study is to present actionable suggestions for the creation of a reliable, efficient and flexible system for text recognition in multilingual

video scenarios, which is synthesized from the literature.

Existing surveys have primarily focused on general scene text detection and recognition, including image-based benchmarks and individual recognition tasks. However, limited attention has been given to the combined analysis of multilingual text detection, script recognition, and video scene environments. This review addresses this gap by systematically analyzing deep learning-based approaches from 2012 to 2024, with emphasis on multilingual capabilities, video-based challenges, datasets, evaluation criteria, and emerging research directions.

The main contribution of this review is the systematic synthesis of deep learning approaches for multilingual text detection and script recognition in video scenes. Unlike previous surveys that primarily focus on general scene text detection or recognition, this study provides an integrated analysis of architectures, multilingual datasets, evaluation metrics, application challenges, and future research trends. The review also establishes a comparative framework to analyze model capabilities, limitations, and research gaps in multilingual video text processing.

The paper begins with an introduction outlining the challenges of multilingual text detection (MTD) and script recognition. The Literature Review summarizes developments in CNNs, RNNs, transformers, and hybrid models. The Methodology section describes the systematic review process and study-selection criteria. The Results and Discussion section presents models, datasets, evaluation metrics, and key challenges, while the Conclusion summarizes the major findings, limitations, and future research directions.

2. Literature Review

In this paper, the latest developments in multilingual text detection and script recognition in video scenes are discussed. It is focussed on the recent deep learning models, architectural improvement and applications as illustrated in figure 1. We developed key techniques: CNNs (EAST, TextBoxes++), RNNs (CRNNs for spatial-temporal modeling), transformers (DETR for dense

scenes), and hybrid models (Mask TextSpotter v3).

To address issues of low-resolution, motion blur, text orientation and occlusion, innovative methods including ESRGAN, deblurring networks, and attention mechanisms are used. This can be used in real-time production of subtitles, translation in augmented reality, assistive technology and document digitization. Thus, this analysis is used to point out some of the strengths and weaknesses of the existing methods as well as to provide directions for future work.

The development of multilingual text detection and script recognition approaches has progressed through different stages during the 2012–2024 period. Early research mainly focused on CNN-based architectures for feature extraction and scene text detection, providing improved visual representation compared with traditional approaches. Later studies introduced RNN-based and CRNN architectures to model sequential text information and improve recognition in dynamic environments. With the advancement of transformer-based architectures, researchers explored global contextual modeling and improved handling of complex text layouts. Recent studies have further investigated hybrid architectures, multimodal learning, transfer learning, and self-supervised approaches to address challenges related to multilingual scenarios, low-resource scripts, and real-time video text processing.

2.1 Deep Learning Techniques in Text Detection in Video Scenes

This section introduces detailed description of the deep-learning-based methods for text detection in video scenes:

Convolutional Neural Networks (CNNs): Convolutional Neural Networks can be used in text detection tasks in video scenes due to their spatial representation capabilities. There are several interesting models, such as EAST (Lu et al., 2021; X. Zhou et al., 2017) and TextBoxes++ (Liao et al., 2018). EAST is a fully convolutional network, which relies on pre-trained backbones such as VGG16 and ResNet and hence is suitable for video surveillance, but is not ideal for small fonts

and complex layouts (Lu et al., 2021; Mayank et al., 2021; Xiang & Luo, 2020; X. Zhou et al., 2017). TextBoxes++ extends the SSD framework to detect oriented and curved text using specialized text-box layers and text-aware non-maximum suppression (NMS), although it can be computationally expensive. (Liao et al., 2018; Lu et al., 2021; Xiang & Luo, 2020; Xue et al., 2021). Each has its unique strengths and weaknesses, as described in Table 1.

Table 1. Comparative Analysis

Feature	EAST	TextBoxes++
Approach	End-to-end detection	Extended SSD framework
Orientation Handling	Limited to rotated bounding boxes	Supports arbitrary orientations
Loss Function	Composite (segmentation + regression)	Regression with specialized anchor boxes
NMS	General-purpose	Text-specific
Real-Time Performance	High	Moderate
Best Use Case	Real-time video text detection	Multi-oriented and densely packed text

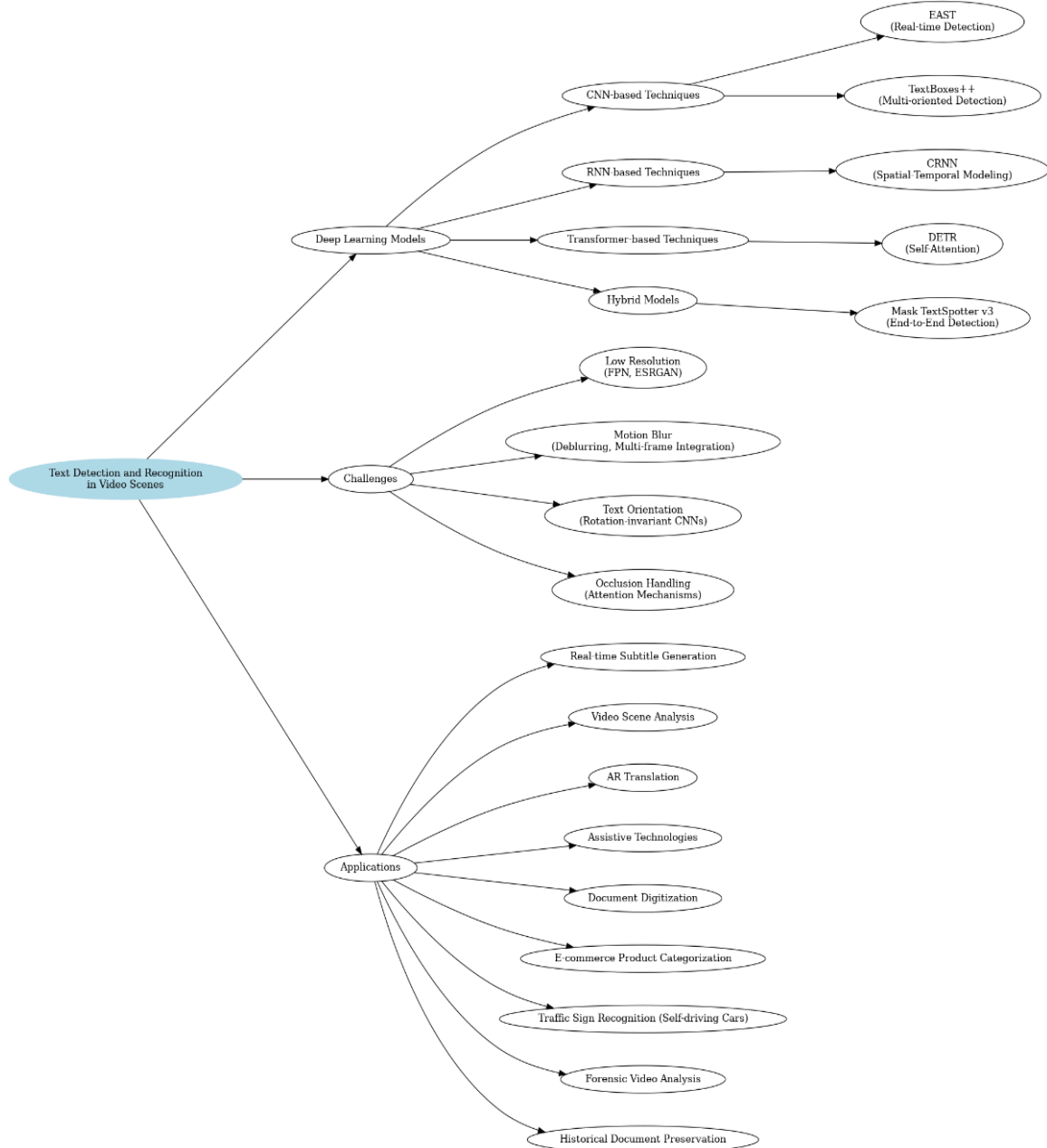


Figure 1. Literature Review Framework for Text Detection and Recognition

Recurrent Neural Networks (RNNs): RNNs, especially LSTMs, are well adapted to model the temporal dependencies existing in video sequences, which benefits text detection tasks in dynamic scenes. CRNN, which combines the CNN with a bidirectional LSTMs structure, is proficient in extracting spatial-temporal features of text and highly suitable for any task such as captions,

security, and live-stream text recognition (H. Li et al., 2017; Y. Wang et al., 2018; Yu et al., 2019). As CRNN achieves high performance by handling challenging scenes like occluded text, it is relatively computationally demanding and needs proper optimization. Its key strengths and challenges are reported in Table 2.

Table 2. Key Strengths and Challenges of CRNNs in Video Text Detection

Aspect	CRNN Approach
Strengths	Combines spatial and temporal modeling for robust text detection.
	Effective for handling occluded, blurred, or sequential text information.
	Bidirectional LSTMs enhance context understanding in both time directions.
Challenges	Computationally intensive due to recurrent layers and multi-frame processing.
	Requires careful optimization for real-time or large-scale applications.
Applications	Automatic subtitle generation, real-time text recognition in videos, and content analysis.

2.2 Transformer-based Models

DETR (DEtection TRansformer) was originally introduced as a transformer-based framework for general object detection, using a CNN backbone, transformer encoder-decoder architecture, and object queries. Its ability to model global relationships has motivated subsequent adaptations for specialized vision tasks, including scene text analysis. In multilingual video text processing, DETR-based approaches require task-specific modifications, such as text-oriented detection heads, temporal modeling strategies, or integration with recognition modules, to address the characteristics of video text scenarios. (Carion et al., 2020; Liang et al., 2022; Yang et al., 2022; Q. Zhou et al., 2023). Without object queries, DETR achieves good performance in terms of oriented and scaled text because of the receptive field of self-attention. DETR-based approaches have shown advantages in

modeling global contextual relationships and handling complex visual layouts. Their effectiveness depends on the dataset, training strategy, and evaluation protocol used in each study. DETR itself was originally designed for single-image object detection and does not inherently perform temporal reasoning across video frames. For video text detection scenarios, DETR-based approaches require additional temporal modules, frame aggregation strategies, or video-specific adaptations to exploit information across consecutive frames. Therefore, improvements in handling occlusion and motion blur should be attributed to these extended architectures rather than the original DETR framework alone. This makes it possible for use in activities such as subtitled video and video content recognition of content. The summary of the strength and challenges of DETR for VT detection are highlighted in table 3.

Table 3. Potential Role and Challenges of DETR-based Architectures in Scene Text and Video Text Analysis

Aspect	DETR Approach
Strengths	Global context modeling using self-attention.
	Robust to variations in text scale, orientation, and layout.
	Effective in dense and cluttered scenes.
	Can provide global contextual modeling through self-attention; temporal reasoning requires additional video-specific extensions.
Challenges	Computationally intensive due to the transformer architecture.
	Requires significant training data for optimal performance in video-specific tasks.
Applications	Potential applications in scene text analysis and video text systems when combined with task-specific adaptations.

Hybrid text spotting approaches, such as Mask TextSpotter v3, integrate detection and recognition components within an end-to-end framework. These models have demonstrated competitive performance in scene text

spotting benchmarks by combining visual feature extraction, contextual modeling, and character recognition. However, performance improvements depend on the dataset characteristics, evaluation metrics, and

experimental settings used in each study. Therefore, comparisons between hybrid models and previous approaches should be interpreted based on reported quantitative evaluations rather than general assumptions of superiority.

The model further has the advantage of allowing segmentation branch accuracy to pixel levels, which enables precise segmentation, especially in more complicated text patterns. Although computationally demanding, Mask TextSpotter v3 demonstrates strong capability in end-to-end scene text spotting by combining text detection, segmentation, and recognition. However, its reported evaluations are primarily based on scene text spotting benchmarks, and its direct applicability to multilingual video environments requires additional adaptation and validation, as shown in Table 4.

Table 4. Key Strengths and Applications of Mask TextSpotter v3

Aspect	Description
Strengths	Combines the precision of segmentation with the contextual understanding of transformers and RNNs.
	Effective in detecting text with varying orientations, scales, and layouts.
	Excels in recognizing text sequences and handling overlapping or cluttered text regions.
Challenges	Increased computational complexity due to the integration of multiple components.
	Requires significant training resources for video-specific tasks.
Applications	Scene text spotting, text recognition, and related visual text analysis tasks. Adaptation may be required for multilingual video-scene applications.

Advanced Methodologies and Models for Overcoming Challenges in Video Text Detection: Such challenges as low resolution, motion blur, and text orientation are solved through methodologies, for example, multi-scale detection in FPN and the use of ESRGAN in low-resolution frames enhancing (ALTUN & TALU, 2019; He et al., 2023; Lin et al., 2017; Shocher et al., 2020; X. Wang, Yu, et al., 2019). For motion blur, deblurring networks along with multi-frame integration solve such cases, whereas CNNs which are rotation invariant handle dynamic scenes, and spatial transformation and attention-related mechanisms can improve robustness to geometric variations and complex visual conditions (Cheon et al., 2022; Finnveden et al., 2021; Zhan & Lu,

2019). CRAFT and other curved text detection approaches have demonstrated effectiveness in handling irregular text shapes. More sophisticated approaches like DB and ABCNet also enhance accuracy for challenging scenarios of orientation variation and different resolutions (Baek et al., 2020; Dueire Lins et al., 2019; Y. Liu et al., 2020, 2021). All these improvements greatly enhance robustness and precision in video text detection.

In multilingual video text analysis, several related but distinct tasks are involved. Text detection refers to locating text regions within video frames, whereas text recognition focuses on converting detected text regions into readable character sequences. Script identification determines the writing system of the detected text, while language identification determines the associated language. Recent vision and language models are often used as supporting components rather than complete video text recognition systems. Their relevance lies in providing visual representations, multilingual language understanding, or cross-modal alignment that can be integrated into end-to-end text detection and recognition frameworks.

2.3 Script Recognition in Multilingual Contexts

Feature Extraction Techniques: Some of the script recognition systems use these CNNs such as ResNet and DenseNet to extract, features from the video frames viably. New techniques like the Advanced CNNs perform well in learning script specific features across the languages it is trained from the stroke level features to script level features CNN-based feature extraction methods have been widely adopted in script identification and scene text recognition because they can learn hierarchical visual representations from text regions. Attention mechanisms further improve feature selection by emphasizing informative spatial regions and character-level features, particularly in complex text environments involving variation in scale, orientation, and background conditions (Xu et al., 2020). It is important to understand that transfer learning approaches have been very useful in the training of CNNs, as the proposed models are fine-tuned from pre-

trained models, and in few-resource language this reached almost state of the art results due to limited annotated data. EfficientNet provides a scalable convolutional feature extraction framework that balances computational complexity and model capacity. However, its effectiveness for multilingual script recognition depends on task-specific training data, script diversity, and evaluation protocols. Therefore, EfficientNet should be considered a general visual backbone rather than a dedicated script recognition method. Feature extraction remains important even today, especially for retaining the feed-forward text character in video recognition systems.

Script recognition and language identification are related but distinct tasks. A script represents the writing system used to represent text, whereas language identification determines the linguistic identity of the text content. Therefore, different languages may share the same script, and some languages may be represented using more than one script depending on historical, regional, or application-specific contexts. In multilingual video text analysis, models and datasets should be classified according to whether they address script identification, language identification, text detection, or text recognition.

Sequence Modeling Approaches: Temporal Information in the Video Scenes is localized using sequence modeling approaches. The bidirectional LSTMs are effective for the temporal property of the text in script matching in frames because they establish context in both the past and the future. Temporal attention mechanisms can emphasize informative frames and support modeling of temporal dependencies in video sequences (Jozefowicz et al., 2015; L. Wang et al., 2018). Here, CNN captures spatial features and RNN models temporal features making CRNNs have the ability to capture spatial temporal features. For better performance, Advanced AC-GRU uses sequence attention maps as learned (Y. Wang et al., 2018; Yu et al., 2019). Transformer-based architectures contribute differently to multilingual video text processing depending on their design. Language models such as

BERT, mBERT, and XLM-RoBERTa provide multilingual contextual representations and are mainly used for language understanding or text-level processing. Vision-based models such as ViT provide visual feature representations, while vision-language models such as ViLBERT and CLIP learn relationships between visual and textual information. These models can support multilingual video text recognition when integrated with detection and recognition pipelines, but they do not independently perform video text detection or recognition. (Khan et al., 2022; Tao et al., 2021).

Multilingual Training Strategies: With the help of multilingual training and with large data sets, script recognition has developed a lot. Even when trained in very different multiple script or multiple language environments such as Wild SynthText and ICDAR MLT result in script invariant features and for Arabic languages and scripts (Macko et al., 2023; Moosa et al., 2023; Tüske et al., 2014). Long datasets have been endeavored to gather in order to improve machine models, including historical scripts and endangered scripts. Application of Multi-task learning can improve multilingual text processing by jointly optimizing related but distinct tasks, including script identification, text recognition, and language classification. These tasks provide complementary information but address different aspects of multilingual video text analysis, and supports low resource languages. Script-aware attention and normalization methods help IET enhance script-specific discriminative feature learning at the same time of keeping IET framework language-generic. To enhance the ability of the models to handle multiple scripts, curriculum learning which trains them in a gradual manner with frequency of scripts ranging from frequent to rare is used. Prior models have also been very useful especially in the enhancement of script recognition systems.

Advancements through Pre-trained Models: The recent improvements in pre-trained models provided a big boost to script recognition through NLP. Even fine-tuned models such as mBERT and XLM-

RoBERTa, where multilingual pre-training is done, have been found to be successful for low resource scripts (Gessler & Zeldes, 2022; Hedderich et al., 2020; Rodrigues da Silva & Caseli, 2021). The Vision-language pre-training model such as ViLBERT and CLIP solves the Vision-Text gap, and improves the multilingual script recognition (Rao et al., 2022; Sun et al., 2019). Contrastive learning, and masked auto-encoding—both self-supervisory approaches, have enhanced performance particularly in low resource languages but identifying novel scripts is still a problem.

Challenges in Zero-shot Script Recognition: One of the major problems with script recognition is the ‘zero-shot recognition’ in which various alphabets within a script not seen during training are recognized during test. However, this is still in progress as pointed out in previous literature. Such meta-learning approaches including in Model-Agnostic Meta-Learning (MAML) have been used to increase the potential of the mode to work on new scripts with less training time (Fernando et al., 2018; Huisman et al., 2021; Yue et al., 2023; Zellers et al., 2022). Matching Networks and Prototypical Networks have been used for few-shot learning of new scripts (Ji et al., 2020; X. Liu et al., 2022). Other techniques, such as Generative models, including GANs, along with better style transfer improves the zero-shot recognition (Diao et al., 2023; X. Liu et al., 2022; Souibgui et al., 2022). The described approach, which concentrates on new scripts, does not contradict the need to recognize various known scripts.

Cross-script Generalization: Solving the problems of building subsets of recognition for scripts involves the ability to read scripts from different writing systems, as well as variations within the major forms of a writing system. As seen, several tags are often written in multiple scripts, and consequently, improving generalization across multiple scripts for recognition of multilingual script is imperative. Thus, the adaptation techniques have been used to reduce the gap between the source and target script domains. Adversarial training methods have attained higher accuracy across scripts by concentrating on

script related landmarks (Bhunja et al., 2019; Cheng et al., 2019; L. Kang et al., 2020). The objective of this paper is to identify generalized forms of writing systems and enhance the stability of the model using methods such as script-blinded stroke analysis and morphing that preserve graph connectivity. Recent development includes the networks for script invariant features while having direct script knowledge. Some of the data augmentation techniques like MixUp and CutMix have been found to be useful in teaching the model script-invariant characteristics (D. Guo et al., 2020; Wu et al., 2021; Yin et al., 2021). Strategies such as gradual fine tuning and layer wise adaptations enhance cross-script transfer learning from high resource scripts to low resource scripts.

As discussed in this section, script recognition in multilingual environments has advanced from simple feature extraction to advanced methods tied to deep learning, transfer learning, and cross-scripting. These improvements have clearly boosted our capability to identify and read text in different writing systems besides having solutions to difficulties such as zero-shot identification as well as domain shift. But successful script identification is just one aspect of script analysis in the general score of video text recognition. The peak of these technologies can be reached only if the technologies used for text detection and script recognition are connected. This integration is significant because they need to build robust systems to deal with problems in natural, multilingual video data. The following section will discuss further how the researchers are making the text detection or the text recognition as an end-to-end application for accumulating and elaborating the textual data in various scenarios of the video scenes.

2.4. Integration of Text Detection and Recognition in Videos

This section provides an in-depth overview of integrating text detection and script recognition for video scene analysis, focusing on end-to-end systems, challenges, and practical applications:

End-to-End Systems / Framework: The most recent approaches in text and script

identification focus more on the joint use of both tasks in End-to-End systems. Some such models usually employ initial two or more convolutional layers for feature extraction and the model is then divided into two streams for detection and recognition sub-tasks. Such a structure enables efficiency in the processing of information and the exchange of information which in turn increase overall performance (H. Li et al., 2017). Furthermore, different tasks integration is applied virtually in all systems with multitask learning to perform detection of text and recognition of script in parallel with possible tasks such as text line segmentation or language identification. This helps models to share parameters and overall performance by generating similar features from different related tasks (Vandenhende et al., 2021). Some of them are used to calculate attention maps for the features needed for detection and recognition purposes. Spatial attention influences the position of text areas and temporal attention addresses sequential content; higher complexities of video scenes are addressed well (Guo et al., 2022).

In addition, recurrent-convolutional architectures to integrate CNN in feature learning with RNN for sequential learning best suited for image sequence (Donahue et al., 2017; K. Kang et al., 2018; Ziat et al., 2017). Transformer based models has also been used for text detection and recognition in video using effective mechanisms for learning long-term dependencies and flexible input sizes (Han et al., 2023; Khan et al., 2022).

Challenges: Accuracy of performance is critical in real time processing, especially for video stream and is difficult when the videos are in high-definition format. Other proposals under consideration to handle this are model pruning, quantization, and hardware acceleration. There are a few pros and cons when it comes to time efficiency: speed and accuracy should be sacrosanct here since in real-life problems there are always sequences where one has to respond quickly and the next problem solved as soon as possible. As more language and script types are added, the issues of scalability become amplified. Further, maintaining a balance between model

complexity and the throughput in the case of different writing systems has been tough. Today's research is to develop models that can handle multiple languages, without needing much additional resources. Sensitivity and specificity are essential, but consistency in detection plus recognition across languages, lithography, or visual conditions is challenging. This is made more difficult by variability in text features such as, font, rotation and compound backgrounds etc. Thus, temporal information in video sequences can be considered the basis of possibilities for the improvement of the detection and the recognition in the stereoscopic application. Adding and switching between languages require different challenges such as values used in new languages or new scripts like low-resource language. The required training strategies that can scale to other writing systems with limited data include open architectures. Complex backgrounds with multiple layers and multiple layers of obstacles like illumination changes, reflections, and occlusion make it challenging to detect texts in terms of segmentation and to recognize them. Furthermore, multi-oriented and curved text processing is a more complex problem and needs more sophisticated models and training. Finally, it makes sense to achieve a balance between model size and performance for better efficiency, especially for deployment on devices or systems with low functionality.

Successful Integrated Systems and Applications: Software platforms for automated generation of video subtitles/subtitles for videotaped presentations in multiple languages/scripts in real time as well as real-time processing of audio and text streams for various types of videos (Bougares & Jouili, 2022). It should also be noted that text detection and recognition technologies are also important for document digitization, more specifically for the identification and conversion of large arrays of documents into machine-readable format, or general archiving, like preservation of historical documents and gaining access to printed media and information search in large archives (Dar & Khan, 2017). AR translation

utilizes real time text to translate the signs, menus, and other texts most frequently encountered by a learner when roaming or engaging in cross-cultural interactions (Kanvinde et al., 2021). Self-driving vehicles apply text recognition for traffic sign and road marking that enhance road safety and Advanced driver-assistance systems (ADAS) (Guerrieri & Parla, 2022). Employment of text detection to prevent obscene content on social media sites improves the safety of the internet for the general public (Drolsbach & Pröllochs, 2024). These models are used in assistive technologies for the visually impaired to read and vocalize environmental text to gain independence (Dhanjal & Singh, 2019). In e-business, these systems are used to enhance product categorization and recommendation mechanisms using videos. It featured in educational applications such as translating educative videos, & creating/editing Quiz application (Taurusta et al., 2019). In cases of forensic examinations, such systems help in determining relevant aspects in the surveillance videos (Blanco-Medina et al., 2020). Last but not least, laying the foundation for identifying text hidden in cultural artefacts and historical videos makes the data more accessible to researchers (Agnolucci et al., 2022; Pustu-Iren et al., 2020). Explorations in real-time processing and high accurate in such systems are demonstrated in real-time applications. Such technologies advance by narrowing the transition between image-based and text-based systems in videos and human computer-interaction systems. The progress in the growth of integrated systems that combine text detection and recognition remains an area that effectively solves actual challenges, ranging from subtitles to assistive systems. Many a times, building and tweaking such systems rely on good quality data sets and evaluation techniques, which is explained in the next section.

3. Method

This study took a structured and systematic approach in investigating the state-of-the-art in multilingual text detection and script recognition in video scene analysis. In order to achieve a comprehensive understanding of

the field and its challenges, each step of the methodology was designed as shown in figure 2.

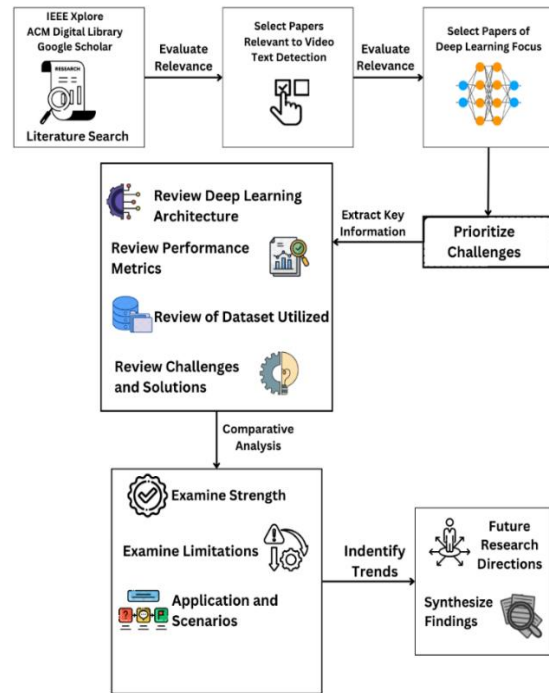


Figure 2. Methodology Steps used for Review Process

3.1 Literature Search and Scope Definition

A systematic literature search was conducted using IEEE Xplore, ACM Digital Library, and Google Scholar to identify studies related to deep learning-based multilingual text detection and script recognition in video scene environments. The search covered publications from January 2012 to December 2024 and was performed using combinations of keywords related to multilingual text recognition, video scene analysis, deep learning architectures, and script identification. The search terms included ("multilingual text detection" OR "scene text recognition" OR "script recognition") AND ("video text" OR "video scene" OR "dynamic text") AND ("deep learning" OR "CNN" OR "RNN" OR "transformer"). Searches were performed in the title, abstract, and keyword fields where supported by the respective databases. The retrieved records were exported, duplicate records were removed, and the remaining studies underwent title/abstract screening followed by full-text eligibility assessment

according to predefined inclusion and exclusion criteria shown in table 5

Table 5. Search Queries Used for Literature Retrieval

Database	Search String
IEEE Xplore	("multilingual text detection" OR "scene text recognition") AND ("deep learning" OR CNN OR RNN OR Transformer) AND ("video scene" OR "video text")
ACM Digital Library	("video text recognition" AND "multilingual script recognition") AND ("deep learning models")
Google Scholar	("deep learning based multilingual text detection" OR "script recognition") AND ("video scenes")

The database searches initially identified 438 records, including 156 records from IEEE Xplore, 102 from the ACM Digital Library, and 180 from Google Scholar. After removing 86 duplicate records, 352 records remained for title and abstract screening. At this stage, 226 records were excluded because they did not meet the predefined scope and eligibility criteria. The remaining 126 full-text articles were assessed for eligibility, of which 68 were excluded: 21 lacked multilingual relevance, 18 focused exclusively on static-image settings without sufficient relevance to video scenes, 14 did not provide quantitative evaluation, 9 lacked sufficient methodological detail, and 6 were review articles or editorials. Finally, 58 studies met the eligibility criteria and were included in the systematic review.

3.2 Criteria for Inclusion and Exclusion

The inclusion and exclusion criteria were defined to ensure that the selected studies were directly relevant deep learning-based multilingual text detection, text recognition, and script identification/recognition tasks in video scene environments. Studies were included if they were published between 2012 and 2024, applied deep learning approaches for multilingual text detection, text recognition, or script recognition, and focused on video scenes, video frames, or dynamic text environments. Studies were also required to provide experimental evaluation using quantitative performance metrics, such as accuracy, precision, recall, F-score, recognition rate, or other relevant evaluation measures.

Studies were excluded if they focused exclusively on static-image-based text recognition without relevance to video scene analysis, addressed only a single language or script without multilingual applicability, lacked experimental validation or quantitative evaluation, were duplicate publications, or did not provide sufficient methodological details for analysis. Review articles, editorials, and studies outside the scope of deep learning-based multilingual text detection and script recognition were also excluded. These criteria ensured that the selected studies provided reliable evidence for evaluating current deep learning approaches in multilingual video scene text analysis.

3.3 Data Extraction and Evaluation

Relevant literature was critically analyzed for extracting relevant information about deep learning models, architectural improvements, and applications. Attention was paid to finding models that support cross-language and cross-script functionalities. The review also included performance metrics, properties of actual datasets employed and restrictions reported to assess these methods for real-world applications.

The techniques identified from the selected studies were categorized based on the reported model architectures, learning strategies, and processing approaches. When a single study employed multiple techniques, each distinct technique was recorded separately. Similar approaches were grouped into broader categories to avoid duplicate representation.

The relationship analysis was constructed by identifying the co-occurrence of techniques reported in the selected studies. Techniques appearing together within the same study were considered related and grouped accordingly. The resulting visualization represents methodological associations extracted from the reviewed literature rather than citation-based or author-based network analysis.

The reviewed studies were compared based on three main dimensions: scalability, robustness, and multilingual capability. Scalability refers to the ability of a method to

handle variations in dataset size, computational requirements, and deployment scenarios. Robustness refers to performance under challenging conditions such as variations in text orientation, image quality, noise, and occlusion. Multilingual capability refers to the ability of a method to process multiple languages or scripts. These dimensions were analyzed based on the qualitative and quantitative evidence reported in the selected studies.

3.4 Comparative Framework for Analysis

A comparative tool to compare the strength and weakness of these identified models was created. Pairwise comparisons were done with regard to scalability, robustness and performance in multilingual context. Further, datasets and methodologies were reviewed concerning their available evaluation to pick up trends, gaps, and overall potential of all approaches.

The conclusions that came out of the analysis was a growing emphasis on a transformer-based model approach. The analysis included the incorporation of multimodal data, the trends for self-supervised learning and generative AI. Some challenges that were reported throughout this study were limited solutions with zero-shot script recognition, difficulties with low-resource languages, and not easy handling at complex video layout. Finally, it was the methodology which was synthesized to give an overall picture of the current situation. The study was focused mainly on effective implementation and their applications in practice, in addition to summarizing the existing issues and future research directions. This research will inspire additional innovation in text detection and script recognition in multilingual text by identifying gaps and the new techniques that are emerging.

4. Findings

This study is presented in this section, along with some of the key critical discussions of multilingual text detection and script recognition in video scene analysis. It comprehensively compares different deep learning models, datasets and methods and presents the studies in terms of performance, use and pros and cons. The discussion starts with comparing the models to demonstrate how the reviewed studies are efficient and have issues using a benchmark-schema. After that, the data sets and benchmarks are compared and reviewed in a substantial study on their similarities, generalization, and applicability to corresponding tasks. Also, all current issues and concerns left unanswered along with potential new research direction are discussed to outline where and how the field requires change and enhancement to progress even more. The study gives a multi-faceted understanding of the advances and shortcomings of studying consumer behavior and helps researchers and practitioners.

4.1 Comparative Analysis of Models

Table 6 shows overviews deep learning methods for text detection in video scenes and compares models by their primary characteristics, strengths, and sources grouped by CNN, RNN, transformers, CNN–RNN hybrids, and more specific categories.

Direct quantitative comparison among the reviewed models is challenging because different studies employ different datasets, annotation strategies, evaluation metrics, and experimental protocols. Therefore, the analysis focuses on reported results within each study rather than establishing a universal ranking among models.

Table 6. Deep Learning Techniques for Text Detection in Video Scenes

Category	Technique/Model	Key Features	Advantages	References
Convolutional Neural Networks (CNNs)	EAST	Fully convolutional network, multi-scale feature merging, task-specific branches for text region and vertex localization.	Efficient and accurate, real-time performance, suitable for multi-oriented text detection.	(Lu <i>et al.</i> , 2021; Xiang & Luo, 2020; X. Zhou <i>et al.</i> , 2017)
	TextBoxes++	Single-shot detector with novel anchor boxes and regression scheme for oriented text, text-specific non-maximum suppression (NMS).	Robust to arbitrary text orientations, efficient single-shot detection, handles closely spaced text instances.	(Liao <i>et al.</i> , 2018; Xue <i>et al.</i> , 2021)

Recurrent Neural Networks (RNNs)	CRNN	Combines CNNs for feature extraction with bidirectional LSTMs for sequence modeling and transcription layers for final label output.	Captures temporal dependencies, effective for dynamic scenes, handles occluded or blurred text across frames.	(H. Li <i>et al.</i> , 2017; Yu <i>et al.</i> , 2019)
Transformer-based Models	DETR	CNN backbone, transformer encoder-decoder architecture, self-attention mechanism, prediction heads for bounding boxes and class predictions.	Captures global context, robust to varying text orientations and scales, effective in complex layouts.	(Carion <i>et al.</i> , 2020; Q. Zhou <i>et al.</i> , 2023)
Hybrid Models	Mask TextSpotter v3 – end-to-end scene text spotting combining text localization and recognition.	CNN backbone, transformer module for context modeling, RNN-based decoder, segmentation branch for precise text localization.	Combines visual feature extraction, contextual modeling, segmentation-based localization, and recognition for end-to-end scene text spotting	(Das <i>et al.</i> , 2024; Park & Paik, 2023)
Low Resolution and Motion Blur	Feature Pyramid Network (FPN)	Multi-scale feature fusion combining low- and high-resolution features for detecting text at varying scales.	Effective for detecting small or distant text, robust to resolution changes in video frames.	(Lin <i>et al.</i> , 2017)
	ESRGAN	Super-resolution GAN to upscale low-resolution frames, enhancing high-frequency details.	Improves detection quality for low-resolution text, realistic high-resolution enhancements.	(He <i>et al.</i> , 2023; X. Wang, Yu, <i>et al.</i> , 2019)
	Deblurring Networks	Preprocessing networks designed to remove motion blur from video frames.	Mitigates motion blur, improves clarity for subsequent text detection stages.	(Shocher <i>et al.</i> , 2020)
Occlusions and Varying Text Orientations	Rotation-invariant CNNs	Polar transformation, circular padding, orientation-aware pooling operations for consistent feature representation.	Handles arbitrarily oriented text, maintains robust detection under diverse text orientations.	(Cheon <i>et al.</i> , 2022; Zhan & Lu, 2019)
	Attention Mechanisms	Spatial, channel, and self-attention mechanisms to focus on relevant image regions and features.	Improves detection in cluttered or occluded scenes, contextualizes text regions across frames.	(Yan <i>et al.</i> , 2021; Xiao <i>et al.</i> , 2022)
Successful Models and Techniques	CRAFT	Character-level detection with linking scores to group text regions, robust to various text shapes and orientations.	Effective for curved text, robust to varying sizes and fonts, handles complex real-world scenarios.	(Baek <i>et al.</i> , 2020)
	DB	Differentiable binarization for end-to-end training, adaptive thresholding, simplified post-processing.	High detection accuracy in varying lighting conditions, efficient end-to-end optimization.	(Liao <i>et al.</i> , 2018)
	ABCNet	Bezier curve representation for curved text, feature pyramid network for multi-scale detection.	Accurate detection of irregular text, robust to perspective distortions and varying orientations.	(Y. Liu <i>et al.</i> , 2020, 2021)

Figure 3 below shows Key Features (blue) and Advantages (green) of different models/techniques. A comparison in terms of counts in Mask R-CNN and DETR

demonstrates that the models' performance, including the features and advantages, is more resistant. Other firms also have average but considerably lower ratios such as FTN and

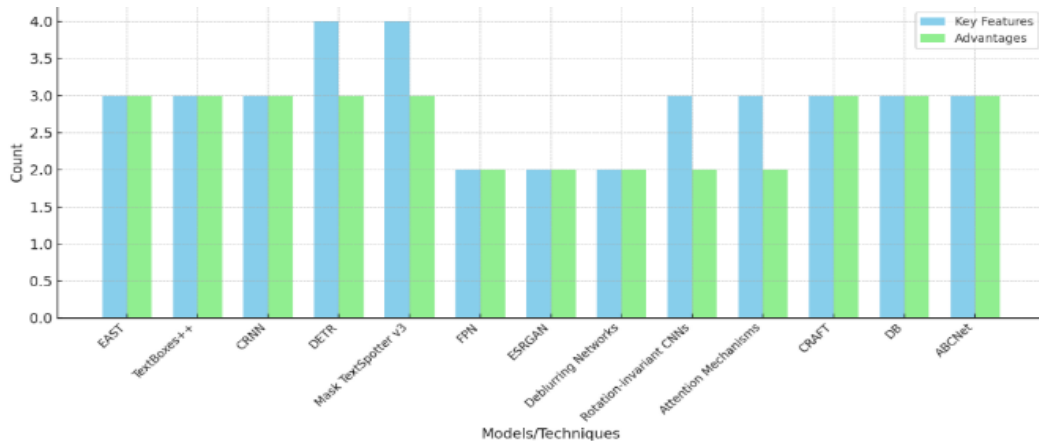


Figure 3. Analysis of Key Features and Advantages of Each Models/Techniques

EASDAM. It focusses on model specific strengths and weaknesses as far as they are manifested in benchmark results for performance or utility.

Pie chart shown in figure 4 helps in pointing out typical and less-frequent problems as well as strategies in the text recognition problems. This explains that successful models (including CNNs, RNNs, and Transformers) mitigate primary concerns, including occlusions, text orientation, low textual resolution, and motion blur. The figure highlights the applicability of various techniques to achieve over these challenges and enhance the performance of the text recognition systems.

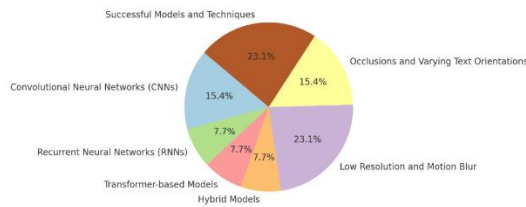


Figure 4. Analysis of Distribution of Techniques by Category

Table 7. Summary of Techniques and Advancements in Multilingual Script Recognition

Category	Key Techniques	Highlights	References
Feature Extraction	CNNs (ResNet, DenseNet), Attention Mechanisms, Transfer Learning, EfficientNet	Hierarchical feature learning; Attention modules focus on script-relevant areas; Pretrained models fine-tuned for low-resource languages; EfficientNet provides a scalable feature-extraction architecture with a balance between model capacity and computational complexity.	(Xu <i>et al.</i> , 2020)
Sequence Modeling	RNNs (LSTM, GRU), CRNN, Attention-based CRNN, Transformers (BERT, ViT)	Combines spatial and temporal modeling; Transformers capture long-range dependencies; Vision Transformers (ViT) offer purely attention-based visual extraction.	(Khan <i>et al.</i> , 2022; Tao <i>et al.</i> , 2021; X. Wang, Yu, <i>et al.</i> , 2019)
Multilingual Training	Large-scale datasets (ICDAR MLT), Multi-task learning, Curriculum learning, Script-aware techniques	Joint training improves generalization; Curriculum learning optimizes for low-resource languages; Script-aware attention enhances dynamic adaptation to various scripts.	(Macko <i>et al.</i> , 2023; Moosa <i>et al.</i> , 2023; Tüske <i>et al.</i> , 2014)
Pre-trained Models	mBERT, XLM-RoBERTa, Vision-Language Models (ViLBERT, CLIP), Self-supervised Learning	Multilingual language models provide contextual text representations, while vision-language models improve visual-text alignment. These models support multilingual text processing when combined with visual detection and recognition frameworks.	(Gessler & Zeldes, 2022; Rao <i>et al.</i> , 2022; Sun <i>et al.</i> , 2019)
Zero-shot Recognition	Meta-learning (MAML), Few-shot learning (Prototypical Networks), Generative approaches (GANs, Style Transfer)	Enables adaptation to unseen scripts with minimal data; GANs generate synthetic examples of unseen scripts; Style transfer supports zero-shot adaptation.	(Diao <i>et al.</i> , 2023; X. Liu <i>et al.</i> , 2022; Yue <i>et al.</i> , 2023)
Cross-script Generalization	Domain adaptation, Script-agnostic feature learning, Data augmentation (MixUp, CutMix), Refined Transfer Learning	Adversarial training ensures robust cross-script features; Augmentation strategies create diverse training data; Layer-wise fine-tuning bridges high-resource and low-resource script domains.	(D. Guo <i>et al.</i> , 2020; K. Kang <i>et al.</i> , 2018; Wu <i>et al.</i> , 2021)

Figure 5 illustrates the distribution of the key techniques across scripts in six categories of script processing. Feature Extraction and Sequence Modeling are on top with five techniques that are implemented, respectively. The following two sections (Pre-

Table 7 discussing about the state of multilingual script recognition, use of feature extraction, sequence modelling, multilingual training, transfer learning, zero shot recognition and cross script generalization, some advanced techniques in transformer, meta learning, GANs, curriculum learning and script independent feature approaches.

Figures 3 and 4 provide a qualitative synthesis of the reviewed studies by categorizing the reported deep learning approaches and their associated characteristics. The categories and model names are extracted from the selected literature, while the presented comparisons indicate the relative presence or discussion frequency of different approaches within the reviewed studies. These figures do not represent direct performance comparisons between models, as the reviewed studies use different datasets, evaluation protocols, and experimental settings.

trained Models, Cross-script Generalization) include four contributions each, followed by Multilingual Training and Zero-shot Recognition, with the latter three sections having three contributions each, highlighting a more targeted approach to various issues.

It also represents the frequency of identified techniques across the reviewed studies and does not indicate the overall contribution or superiority of any approach.

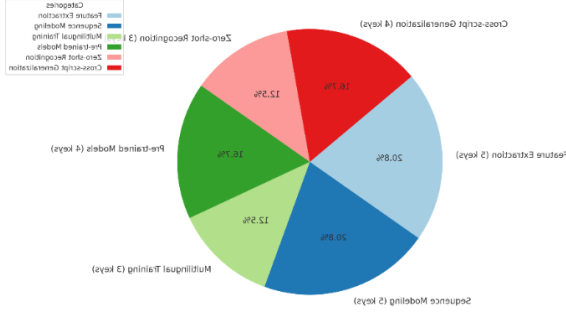


Figure 5. Representation of Number of Key Distribution by Category

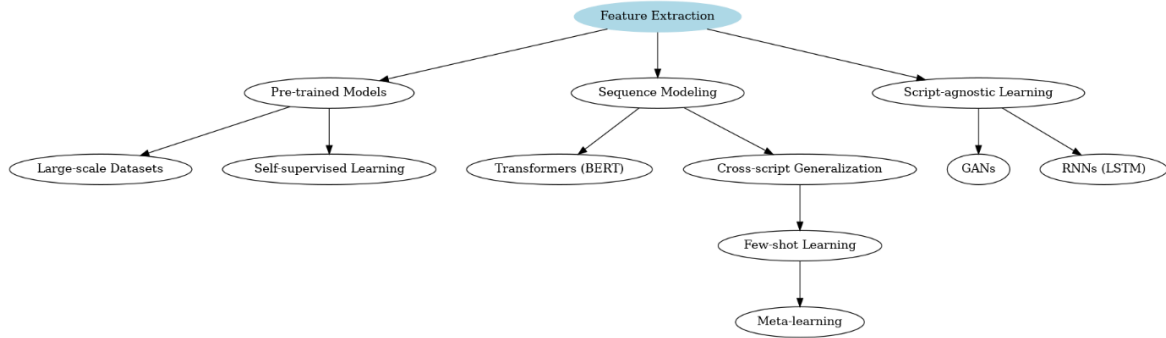


Figure 6. Representation of Relationships Between Categories and Techniques

Figure 6 illustrates the methodological relationships among techniques identified from the reviewed studies based on their reported combinations and applications. It also illustrates the relationship and interaction among the various techniques, within the overall categories of methods. The important connections are indicated as an interaction between Pre-trained Models and Feature Extraction, Feature Extraction and Sequence Modeling. There are several such applications that span across multiple areas, and show significant centrality for progress in research, examples of such applications are Transfer Learning and Transformers.

4.2 Analysis of Datasets and Evaluation Metrics

In this review, multilingual coverage refers to datasets or approaches that support multiple languages or writing systems rather than a single language/script. Resources containing only one language or script are considered monolingual benchmarks and are included only when they provide supporting evidence for specific tasks such as text

detection, recognition, or script identification. Datasets covering multiple scripts or languages are classified as multilingual resources based on their reported annotation and evaluation settings.

Table 8 presents an overview of datasets used in multilingual text detection and recognition research, including their characteristics, size, advantages, limitations, applications, and challenges. The datasets are categorized based on their relevance to video scene text analysis. Video-based datasets are considered primary resources, while image-based benchmark datasets are included only

when they contribute to model development through pre-training, transfer learning, or evaluation of specific components of video text recognition systems. This classification ensures consistency with the review scope while highlighting the contribution of widely used benchmark datasets.

The datasets are categorized based on their original data format and role in the reviewed studies. Video datasets containing temporal information are considered primary resources, whereas image-based datasets are treated as supporting benchmarks for specific components such as detection, recognition, or script identification.

The datasets included in this review are categorized according to their primary task contribution, including text detection, text recognition, script identification, or multilingual evaluation, rather than assuming that all multilingual datasets directly perform language identification.

Table 8. Overview of Publicly Available Datasets for Text Detection and Recognition

Dataset	Key Features	Size	Advantages	Limitations	Applications / Role in Study	Key Challenges	References
YouTube Video Text (YVT)	Real-world video sequences containing text; provides temporal information	~30 hours of video	Real-world video scenarios; temporal continuity; suitable for tracking	Limited language diversity; predominantly English content	Primary video dataset for text detection, recognition, and tracking	Motion blur, compression artifacts, illumination variation, occlusion, temporal instability	Ajwani & Arolkar (2019); Dong <i>et al.</i> (2022); YouTube Video Text (2024)
ICDAR MLT	Multilingual scene-text dataset covering multiple scripts and languages	>20,000 images/frames	Broad multilingual coverage; diverse real-world scenes	Mainly image-oriented; limited temporal information	Supporting multilingual benchmark for transfer and cross-dataset evaluation	Script diversity, writing direction, character variation	Nayef <i>et al.</i> (2019); Sharma <i>et al.</i> (2015); Yan <i>et al.</i> (2023)
COCO-Text	Natural-scene text derived from MS COCO	63,686 images; 173,589 text instances	Diverse scenes; rich annotations	Mainly static imagery; limited multilingual/video content	Supporting image-based benchmark	Cluttered backgrounds, small text, occlusion, orientation	Veit <i>et al.</i> (2016); COCO-Text (2023)
MSRA-TD500	Multi-oriented English and Chinese text	500 images	Strong benchmark for oriented text detection	Small dataset; no video	Supporting orientation-robust text detection benchmark	Arbitrary orientation, long text lines, scale variation	Yao <i>et al.</i> (2012)
CTW	Chinese text in natural scenes with character-level annotations	32,285 images; ~1,018,402 characters	Large-scale Chinese character coverage	Primarily Chinese; static images	Supporting script-specific benchmark	Dense text, small characters, font and scene variation	Liu <i>et al.</i> (2019); Yuan <i>et al.</i> (2019)
CVSI (IIIT-ILST)	Indian-script scene text including Devanagari, Bengali, and Telugu	~1,000 images/script	Represents less-studied Indic scripts	Static images; limited scripts	Supporting Indic-script recognition benchmark	Script-specific characteristics and visually similar characters	Mathew <i>et al.</i> (2017); Sharma <i>et al.</i> (2015)
SynthText in the Wild	Synthetic natural-scene text with variations in font, size, orientation, and placement	~800,000 images; ~8 million text instances	Large scale; useful for pre-training and augmentation	Synthetic-to-real domain gap; limited temporal realism	Synthetic pre-training/data-augmentation resource	Domain shift between synthetic and real-world text	Gupta <i>et al.</i> (2016); Sagduyu <i>et al.</i> (2018)

Real-time performance reported in the reviewed studies is interpreted based on the experimental conditions provided by each study, including hardware configuration, input resolution, batch size, dataset, and processing pipeline. Since these conditions vary across studies, reported FPS or latency values cannot be directly compared unless evaluated under comparable settings.

Figure 7 presents the qualitative contribution and diversity of datasets used in multilingual text detection and recognition studies. The reviewed datasets differ in terms of data format (synthetic images, still images, and video sequences), language/script coverage, and annotation characteristics. Therefore, direct percentage comparison is not appropriate. The figure provides a qualitative overview of dataset diversity, their

role in the review, and their contribution to multilingual text detection and recognition research.

Dataset	Data Format	Language/Script Coverage	Role in Review
 SynthText in the Wild	Synthetic images	Multilingual synthetic text	Supporting pre-training / synthetic benchmark.
 ICDAR MLT	Still images	Multilingual scene text	Supporting multilingual benchmark.
 CVSI (IIIT-ILST)	Still images	Indic scripts	Supporting script recognition benchmark.
 CTW	Still images	Chinese text	Supporting script-specific benchmark.
 MSRA-TD500	Still images	English and Chinese	Supporting orientation/text detection benchmark.
 YouTube Video Text (YVT)	Video sequences	Video scene text	Primary video dataset.
 COCO-Text	Still images	Primarily English scene text	Supporting image-based benchmark.

Datasets are compared qualitatively because they differ in units, coverage, and annotation structure.

Figure 7. Qualitative Dataset Contribution and Coverage in Multilingual Text Detection and Recognition Research

The discussion of computational efficiency has been revised to clarify that real-time performance depends on reported experimental conditions, including hardware

configuration, input resolution, and evaluation settings. The revised manuscript now considers FPS, latency, and computational requirements only when these values are explicitly reported in the reviewed studies.

Table 9 below presents a list of evaluation metrics for text detection and recognition system where we provide definitions, calculation, relevance, use, and citation for each of them. They offer a clear framework by which the performance of such systems can be evaluated and contrasted efficiently.

The reviewed studies use different evaluation metrics depending on the addressed task. Detection-oriented approaches commonly report localization metrics such as IoU, precision, recall, F-score, and mAP, while recognition-oriented approaches use sequence-level measures such as CER, WER, and recognition accuracy. Computational efficiency is evaluated using measures such as FPS, latency, and resource requirements when reported.

4.3 Open Challenges and Future Directions

The table 10 below summarizes key challenges, implications, research needs, and potential solutions in advancing video text recognition across diverse scenarios. It highlights cutting-edge areas like zero-shot script recognition, domain adaptation, handling complex layouts, real-time processing, video quality robustness, and multilingual code-switching. These advancements are pivotal for future applications in real-world contexts.

Table 11 summarizes proposed and emerging research directions for multilingual video text recognition. Some directions, including privacy-preserving recognition, fairness, explainable AI, and NLP integration, are prospective areas that require further task-specific validation rather than established

findings from the reviewed studies. Generative AI is discussed as a future opportunity for improving data availability, restoration, and low-resource script handling; however, its effectiveness for multilingual video text recognition requires further investigation.

5. Conclusion

This review presents new trends, limitations, and possible development in the field of multilingual text detection and script recognition for video scene analysis. Convolutional neural networks, recurrent neural networks, and transformer-based approaches have demonstrated strong capability in multilingual text detection and recognition by learning more effective visual and contextual representations compared with many traditional approaches. However, the extent of improvement depends on the datasets, evaluation protocols, and experimental settings reported in individual studies. Integrated text detection and recognition systems can provide end-to-end processing advantages by combining localization and recognition capabilities. However, performance and computational efficiency depend on model architecture, dataset characteristics, and deployment requirements. Some of the outstanding problems are zero resource script identification, domain adaptability problems, curved text/script recognition, and real time, high-definition video processing on low end portable devices. The reviewed literature indicates potential applications of multilingual text detection and recognition in video accessibility, document analysis, content understanding, translation support, and intelligent transportation systems, although their effectiveness depends on the task, dataset, and deployment conditions.

Table 9. Commonly Used Evaluation Metrics for Text Detection and Recognition

Metric	Definition	Formula	Importance	Applications	References
Precision	Ratio of correctly detected text instances to total detected instances	$\text{Precision} = \text{TP} / (\text{TP} + \text{FP})$	Measures the accuracy of positive predictions	Evaluating detection accuracy	(Mathew <i>et al.</i> , 2017)
Recall	Ratio of correctly detected text instances to total actual text instances	$\text{Recall} = \text{TP} / (\text{TP} + \text{FN})$	Measures the completeness of positive predictions	Ensuring comprehensive detection	(Dutta <i>et al.</i> , 2018)

F1-Score	Harmonic mean of precision and recall	$F1 = 2 * (\text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall})$	Balances precision and recall	Overall performance evaluation of detection systems	(Dutta <i>et al.</i> , 2018)
Intersection over Union (IoU)	Measures overlap between predicted and ground truth bounding boxes	$\text{IoU} = \text{Area of Overlap} / \text{Area of Union}$	Evaluates localization accuracy	Object detection and segmentation tasks	(Everingham <i>et al.</i> , 2014)
Mean Average Precision (mAP)	Mean Average Precision (mAP) summarizes detection performance by averaging Average Precision (AP) values across classes or evaluation categories. The calculation depends on the benchmark protocol, including the IoU threshold used to determine correct detections. Different datasets may adopt different IoU settings, such as mAP@0.5 or COCO-style averaged IoU thresholds.	$\text{mAP} = (1/N) \sum \text{AP}_i$	Evaluates detection performance across multiple classes	Comparing models with different precision-recall trade-offs	(Everingham <i>et al.</i> , 2014)
Character Error Rate (CER)	Edit distance between recognized text and ground truth at character level	$\text{CER} = (S + D + I) / N$	Fine-grained evaluation of text recognition accuracy	Optical Character Recognition (OCR) systems	(Baek <i>et al.</i> , 2020)
Word Error Rate (WER)	Measures recognition errors by comparing the recognized word sequence with the reference transcription using insertion, deletion, and substitution errors.	$\text{WER} = (S + D + I) / N$	Evaluation of word-level recognition accuracy by measuring transcription differences between predicted and reference text.	Speech recognition, machine translation, and text recognition systems	(Morris <i>et al.</i> , 2004)
Computational Efficiency	Metrics include processing time per frame, FPS (Frames Per Second)	N/A	Useful for assessing computational performance and hardware-related efficiency	Evaluating model performance on different hardware configurations	(Georgescu <i>et al.</i> , 2021)

Table10. Unresolved Challenges, Implications and Research Needs

Aspect	Challenge	Implications	Research Needs	Potential Solutions
Zero-shot Script Recognition	Recognizing unseen scripts and languages during training.	Limits extensibility to new languages and scripts.	Developing models for new scripts based on acquired features and linguistic typology.	Meta-learning, transfer learning, and using linguistic typology for model architecture.
Domain Adaptation	Models struggle with transferring to new domains or unseen video types.	Slower, less effective in real-world scenarios with varied videos.	Hardware-efficient generalization strategies requiring minimal parameter tuning.	Adversarial training, domain-invariant feature transfer, and few-shot learning.
Handling Complex Layouts	Difficulty reading non-standard placements, curved texts, or artistic styles.	Limited effectiveness in cases with innovative text formats.	Models capable of deriving meaning from layout and style information.	Graph-based representations, multi-scale analysis, and layout-embedded attention techniques.
Real-time High-res Processing	Achieving timeliness with fidelity in high-resolution video streams.	Limits applications requiring rapid responses, e.g., autonomous vehicles.	Enhanced hardware designs and less complex, efficient algorithms.	Edge computing accelerators, device-specific designs, and universal performance benchmarks.
Robustness to Video Quality	Issues with motion blur, compression artifacts, and poor lighting.	Reduced reliability in varied environmental video capture conditions.	Techniques to improve model resistance to quality issues and adaptive preprocessing.	Simulating quality degradations during training and real-time video enhancement techniques.
Multilingual Code-switching	Identifying text in videos switching between languages or scripts.	Poor performance in mixed-language scenarios, limiting broader usability.	Models that handle multilingual frames or sequences effectively.	Language identification integration, co-ranking of languages, and developing diverse multilingual datasets.

Table 11. Emerging Approaches and Research Directions in Video Text Recognition

Research Area	Approach	Potential Benefits	Research Focus	Challenges	Application Areas
Leveraging Multimodal Data	Adding text detection and recognition to other	Faster analysis and better scene understanding.	Developing unified architectures to	Misalignment between modalities and	Object/scene detection, closed-captioning, video accessibility.

	applied modalities such as audio and video context.		integrate multiple modalities.	sound fusion of dissimilar data forms.	
Incorporating Generative AI	Applying generative models to improve training datasets or text recognition.	Potential support for low-resource languages, rare scripts, and text restoration through future research.	Using few-shot learning and style transfer for script recognition.	Ethical concerns regarding synthetic data usage and bias introduction.	Possible applications in text restoration and data augmentation, requiring further validation in multilingual video text recognition scenarios.
Improving Real-time Processing	Releasing efficient architectures and chip optimizations.	Real-time applicability on more devices.	Low-power pruning, model quantization, and hardware-aware NAS.	Balancing computation efficiency and accuracy.	Distributed processing in edge computing, real-time analytics.
Cross-lingual Transfer Learning	Creating models for knowledge transfer across languages and scripts.	Improved performance with minimal additional training.	Exploring meta-learning, universal language models, and linguistic typology integration.	Lack of evaluation frameworks for cross-lingual transfer.	Translation and script recognition in low-resource settings.
Advancing Contextual Understanding	Adding temporal and semantic context to video text recognition.	Improved comprehension of text messages in scenes.	Scene parsing, video understanding models, temporal modeling, and semantic integration.	Incorporating multi-frame information into recognition models.	Text recognition in augmented reality and contextual video analysis.
Exploring Self-supervised Learning	Utilizing unlabelled video data for pre-training.	Enhanced generalization and reduced reliance on labelled data.	Contrastive learning and multi-task pretraining for text representation.	Designing effective self-supervised tasks.	Scalable video AI models for text recognition.
Addressing Ethical Concerns and Biases	Reducing bias in multilingual text recognition.	Equal performance across languages and cultures.	Bias identification, fairness metrics development, and inclusive collaboration with linguists.	Managing fairness in diverse linguistic contexts.	Multilingual text recognition systems with equitable performance.
Advancing Explainable AI	Developing interpretable models and methods for text detection and recognition.	Increased trust and control in recognition systems.	Attention visualization, saliency maps, and explainable neural text recognition.	Aligning model transparency with regulatory needs.	Error identification in text recognition systems for complex applications.
Integrating Text Recognition with NLP	Combining text recognition with text analysis tasks.	Enhanced comprehension of text content and context.	Aligning text recognition with tasks like named entity recognition, sentiment analysis, and machine translation.	Seamless integration with large language models.	Automated censorship, video tagging, and summarization.
Privacy-Preserving Text Recognition	Developing privacy-aware text detection and recognition algorithms.	Ensures compliance with data protection regulations.	Federated learning, differential privacy, and secure multi-party computation for distributed text recognition.	Balancing privacy preservation with performance optimization.	Privacy-conscious text analytics in sensitive domains.

Future studies may investigate privacy-aware recognition, fairness, explainable AI, and integration with language-understanding models; however, these areas remain prospective directions and require further empirical validation specifically for multilingual video text recognition. These technologies also show potential in translation, augmented reality, education, and cultural document analysis, although further validation is required across diverse real-world deployment settings. The potential impact and future development opportunities of these technologies extend across multiple application domains.

References

- Agnolucci, L., Galteri, L., Bertini, M., & Del Bimbo, A. (2022). Restoration of analog videos using Swin-UNet. *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 6985–6987. <https://doi.org/10.1145/3503161.3547730>
- Ajwani, P., & Arolkar, H. (2019). Tubeextractor: a crawler and converter for generating research dataset from YouTube videos. *International Journal of Computer Applications*, 182(50), 14–17. <https://doi.org/10.5120/ijca2019918738>
- Altun, S., & TALU, M. F. (2019). Aynı şartlar altında farklı üretici çekimeli ağların karşılaştırılması. *2019 3rd International Symposium on Multidisciplinary Studies*

- and Innovative Technologies (ISMSIT), pp. 1–6.
<https://doi.org/10.1109/ISMSIT.2019.8932786>
- Baek, Y., Nam, D., Park, S., Lee, J., Shin, S., Baek, J., Lee, C. Y., & Lee, H. (2020). Clevel: character-level evaluation for text detection and recognition tasks. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2404–2412.
<https://doi.org/10.1109/CVPRW50498.2020.00290>
- Bhunia, A. K., Das, A., Bhunia, A. K., Kishore, P. S. R., & Roy, P. P. (2019). Handwriting recognition in low-resource scripts using adversarial learning. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4762–4771.
<https://doi.org/10.1109/CVPR.2019.00490>
- Blanco-Medina, P., Fidalgo, E., Alegre, E., Alaiz-Rodríguez, R., Jáñez-Martino, F., & Bonnici, A. (2020). Rectification and super-resolution enhancements for forensic text recognition. *Sensors*, 20(20), 5850.
<https://doi.org/10.3390/s20205850>
- Bougares, F., & Jouili, S. (2022). End-to-End speech translation of Arabic to english broadcast news. *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, pp. 312–319.
<https://doi.org/10.18653/v1/2022.wanlp-1.29>
- Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-End object detection with Transformers (pp. 213–229). https://doi.org/10.1007/978-3-030-58452-8_13
- Cheng, C., Huang, Q., Bai, X., Feng, B., & Liu, W. (2019). Patch aggregator for scene text script identification. 2019 *International Conference on Document Analysis and Recognition (ICDAR)*, pp. 1077–1083.
<https://doi.org/10.1109/ICDAR.2019.00175>
- Cheon, J., Baek, S., & Paik, S.-B. (2022). Invariance of object detection in untrained deep neural networks. *Frontiers in Computational Neuroscience*, 16.
<https://doi.org/10.3389/fncom.2022.1030707>
- Dar, W. T., & Khan, M. N. A. (2017). Click-Free, video-based document capture - methodology and evaluation. 2017 *14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, pp. 21–26.
<https://doi.org/10.1109/ICDAR.2017.344>
- Das, A., Biswas, S., Pal, U., & Lladós, J. (2024). Diving into the depths of spotting text in multi-domain noisy scenes. 2024 *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 410–417.
<https://doi.org/10.1109/ICRA57147.2024.10611120>
- Dhanjal, A. S., & Singh, W. (2019). Tools and techniques of assistive technology for hearing impaired people. 2019 *International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, pp. 205–210.
<https://doi.org/10.1109/COMITCon.2019.8862454>
- Diao, X., Shi, D., Li, J., Shi, L., Yue, M., Qi, R., Li, C., & Xu, H. (2023). Toward zero-shot character recognition: a gold standard dataset with radical-level annotations. *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 6869–6877.
<https://doi.org/10.1145/3581783.3612201>
- Donahue, J., Hendricks, L. A., Rohrbach, M., Venugopalan, S., Guadarrama, S., Saenko, K., & Darrell, T. (2017). Long-Term recurrent convolutional networks for visual recognition and description. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4), 677–691.
<https://doi.org/10.1109/TPAMI.2016.2599174>
- Dong, J., Wang, Y., Chen, X., Qu, X., Li, X., He, Y., & Wang, X. (2022). Reading-Strategy inspired visual representation learning for text-to-video retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(8), 5680–5694.
<https://doi.org/10.1109/TCSVT.2022.3150959>
- Drolsbach, C. P., & Pröllochs, N. (2024). Content moderation on social media in the EU: insights from the DSA transparency database. *Companion Proceedings of the ACM Web Conference 2024*, pp. 939–942.
<https://doi.org/10.1145/3589335.3651482>
- Dueire Lins, R., Bernardino, R., & Jesus, D. M. (2019). A quality and time assessment of binarization algorithms. 2019 *International Conference on Document Analysis and Recognition (ICDAR)*, pp. 1444–1450.
<https://doi.org/10.1109/ICDAR.2019.00232>
- Dutta, K., Krishnan, P., Mathew, M., & Jawahar, C. V. (2018). Towards spotting and recognition of handwritten words in Indic

- scripts. *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pp. 32–37.
<https://doi.org/10.1109/ICFHR-2018.2018.00015>
- Everingham, M., Eslami, S. M. A., Van Gool, L., Williams, C. K. I., Winn, J., & Zisserman, A. (2014). The pascal visual object classes challenge: a retrospective. *International Journal of Computer Vision*, 111(1), 98–136. <https://doi.org/10.1007/s11263-014-0733-5>
- Fernando, C., Sygnowski, J., Osindero, S., Wang, J., Schaul, T., Teplyashin, D., Sprechmann, P., Pritzel, A., & Rusu, A. (2018). Meta-learning by the baldwin effect. *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, pp. 109–110.
<https://doi.org/10.1145/3205651.3205763>
- Finnveden, L., Jansson, Y., & Lindeberg, T. (2021). Understanding when spatial transformer networks do not support invariance, and what to do about it. *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 3427–3434.
<https://doi.org/10.1109/ICPR48806.2021.9412997>
- Georgescu, A.-L., Pappalardo, A., Cucu, H., & Blott, M. (2021). Performance vs. hardware requirements in state-of-the-art automatic speech recognition. *EURASIP Journal on Audio, Speech, and Music Processing*, 2021(1), 28.
<https://doi.org/10.1186/s13636-021-00217-4>
- Gessler, L., & Zeldes, A. (2022). Microbert: effective training of low-resource monolingual berts through parameter reduction and multitask learning. *Proceedings of the 2nd Workshop on Multi-Lingual Representation Learning (MRL)*, pp. 86–99.
<https://doi.org/10.18653/v1/2022.mrl-1.9>
- Guerrieri, M., & Parla, G. (2022). Smart tramway systems for smart cities: a deep learning application in ADAS systems. *International Journal of Intelligent Transportation Systems Research*, 20(3), 745–758. <https://doi.org/10.1007/s13177-022-00322-4>
- Guo, D., Kim, Y., & Rush, A. (2020). Sequence-Level mixed sample data augmentation. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5547–5552. <https://doi.org/10.18653/v1/2020.emnlp-main.447>
- Guo, M.H., Xu, T.X., Liu, J.J., Liu, Z.N., Jiang, P.T., Mu, T.J., Zhang, S.H., Martin, R. R., Cheng, M.-M., & Hu, S.-M. (2022). Attention mechanisms in computer vision: a survey. *Computational Visual Media*, 8(3), 331–368.
<https://doi.org/10.1007/s41095-022-0271-y>
- Guo, Z., Zhao, Z., Jin, W., Wang, D., Liu, R., & Yu, J. (2022). Taohighlight: commodity-aware multi-modal video highlight detection in e-commerce. *IEEE Transactions on Multimedia*, 24, 2606–2616.
<https://doi.org/10.1109/TMM.2021.3087001>
- Gupta, A., Vedaldi, A., & Zisserman, A. (2016). Synthetic data for text localisation in natural images. **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**, pp. 2315–2324.
<https://doi.org/10.1109/CVPR.2016.254>
- Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., Tang, Y., Xiao, A., Xu, C., Xu, Y., Yang, Z., Zhang, Y., & Tao, D. (2023). A survey on vision Transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1), 87–110.
<https://doi.org/10.1109/TPAMI.2022.3152247>
- He, X., Luo, Z., Li, Q., Chen, H., & Li, F. (2023). DG-GAN: a high quality defect image generation method for defect detection. *Sensors*, 23(13), 5922.
<https://doi.org/10.3390/s23135922>
- Hedderich, M. A., Adelani, D., Zhu, D., Alabi, J., Markus, U., & Klakow, D. (2020). Transfer learning and distant supervision for multilingual Transformer models: a study on african languages. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2580–2591.
<https://doi.org/10.18653/v1/2020.emnlp-main.204>
- Huisman, M., van Rijn, J. N., & Plaat, A. (2021). A survey of deep meta-learning. *Artificial Intelligence Review*, 54(6), 4483–4541.
<https://doi.org/10.1007/s10462-021-10004-4>
- Ji, Z., Chai, X., Yu, Y., Pang, Y., & Zhang, Z. (2020). Improved prototypical networks for few-shot learning. *Pattern Recognition Letters*, 140, 81–87.
<https://doi.org/10.1016/j.patrec.2020.07.015>

- Jozefowicz, R., Zaremba, W., & Sutskever, I. (2015). An empirical exploration of recurrent network architectures. *Proceedings of the 32nd International Conference on Machine Learning*, 37, pp. 2342–2350.
- Kang, K., Li, H., Yan, J., Zeng, X., Yang, B., Xiao, T., Zhang, C., Wang, Z., Wang, R., Wang, X., & Ouyang, W. (2018). T-CNN: tubelets with convolutional neural networks for object detection from videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(10), 2896–2907. <https://doi.org/10.1109/TCSVT.2017.2736553>
- Kang, L., Rusinol, M., Fornes, A., Riba, P., & Villegas, M. (2020). Unsupervised adaptation for synthetic-to-real handwritten word recognition. *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 3491–3500. <https://doi.org/10.1109/WACV45572.2020.9093392>
- Kanvinde, A., Revadekar, A., Tamse, M., Kalbande, D. R., & Bakereyala, N. (2021). Bidirectional sign language translation. *2021 International Conference on Communication Information and Computing Technology (ICCICT)*, pp. 1–5. <https://doi.org/10.1109/ICCICT50803.2021.9510146>
- Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., & Shah, M. (2022). Transformers in vision: a survey. *ACM Computing Surveys*, 54(10s), 1–41. <https://doi.org/10.1145/3505244>
- Li, H., Wang, P., & Shen, C. (2017). Towards end-to-end text spotting with convolutional recurrent neural networks. *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 5248–5256. <https://doi.org/10.1109/ICCV.2017.560>
- Li, J., Li, Y., Xiang, X., Xia, S.-T., Dong, S., & Cai, Y. (2020). Tnt: an interpretable tree-network-tree learning framework using knowledge distillation. *Entropy*, 22(11), 1203. <https://doi.org/10.3390/e22111203>
- Liang, T., Bao, H., Pan, W., & Pan, F. (2022). Traffic sign detection via improved sparse R-CNN for autonomous vehicles. *Journal of Advanced Transportation*, 2022, 1–16. <https://doi.org/10.1155/2022/3825532>
- Liao, M., Shi, B., & Bai, X. (2018). TextBoxes++: a single-shot oriented scene text detector. *IEEE Transactions on Image Processing*, 27(8), 3676–3690. <https://doi.org/10.1109/TIP.2018.2825107>
- Lin, T.-Y., Dollar, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature pyramid networks for object detection. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 936–944. <https://doi.org/10.1109/CVPR.2017.106>
- Liu, X., Gao, W., Li, R., Xiong, Y., Tang, X., & Chen, S. (2022). One shot ancient character recognition with siamese similarity network. *Scientific Reports*, 12(1), 14820. <https://doi.org/10.1038/s41598-022-18986-z>
- Liu, X., Meng, G., & Pan, C. (2019). Scene text detection and recognition with advances in deep learning: a survey. *International Journal on Document Analysis and Recognition*, (IJ DAR), 22(2), 143–162. <https://doi.org/10.1007/s10032-019-00320-5>
- Liu, Y., Chen, H., Shen, C., He, T., Jin, L., & Wang, L. (2020). Abcnet: real-time scene text spotting with adaptive bezier-curve network. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9806–9815. <https://doi.org/10.1109/CVPR42600.2020.00983>
- Liu, Y., Jin, L., Zhang, S., Luo, C., & Zhang, S. (2019). Curved scene text detection via transverse and longitudinal sequence connection. *Pattern Recognition*, 90, pp. 337–345. <https://doi.org/10.1016/j.patcog.2019.02.002>
- Liu, Y., Shen, C., Jin, L., He, T., Chen, P., Liu, C., & Chen, H. (2021). Abcnet v2: adaptive bezier-curve network for real-time end-to-end text spotting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–1. <https://doi.org/10.1109/TPAMI.2021.3107437>
- Lu, M., Mou, Y., Chen, C.-L., & Tang, Q. (2021). An efficient text detection model for street signs. *Applied Sciences*, 11(13), 5962. <https://doi.org/10.3390/app11135962>
- Macko, D., Moro, R., Uchendu, A., Lucas, J., Yamashita, M., Pikuliak, M., Srba, I., Le, T., Lee, D., Simko, J., & Bielikova, M. (2023). Multitude: large-scale multilingual machine-generated text detection benchmark. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9960–9987.

- <https://doi.org/10.18653/v1/2023.emnlp-main.616>
- Mathew, M., Jain, M., & Jawahar, C. V. (2017). Benchmarking scene text recognition in Devanagari, Telugu and Malayalam. *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, pp. 42–46. <https://doi.org/10.1109/ICDAR.2017.364>
- Mayank, Bhowmick, S., Kotecha, D., & Rege, P. P. (2021). Natural scene text detection using deep neural networks. *2021 6th International Conference for Convergence in Technology (I2CT)*, pp. 1–6. <https://doi.org/10.1109/I2CT51068.2021.9418116>
- Mehrish, A., Majumder, N., Bharadwaj, R., Mihalcea, R., & Poria, S. (2023). A review of deep learning techniques for speech processing. *Information Fusion*, 99, 101869. <https://doi.org/10.1016/j.inffus.2023.101869>
- Merino-Gracia, C., & Mirmehdi, M. (2014). Real-time text tracking in natural scenes. *IET Computer Vision*, 8(6), 670–681. <https://doi.org/10.1049/iet-cvi.2013.0217>
- Mirza, A., Zeshan, O., Atif, M., & Siddiqi, I. (2020). Detection and recognition of cursive text from video frames. *EURASIP Journal on Image and Video Processing*, 2020(1), 34. <https://doi.org/10.1186/s13640-020-00523-5>
- Moeslund, T. B., Thomas, G., Hilton, A., Carr, P., & Essa, I. (2017). Computer vision in sports. *Computer Vision and Image Understanding*, 159, 1–2. <https://doi.org/10.1016/j.cviu.2017.05.006>
- Moosa, I. M., Akhter, M. E., & Habib, A. B. (2023). Does transliteration help multilingual language modeling? findings of the association for computational linguistics: EACL 2023, 670–685. <https://doi.org/10.18653/v1/2023.findings-eacl.50>
- Morris, A. C., Maier, V., & Green, P. D. (2004). From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition. *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, pp. 2765–2768.
- Nayef, N., Patel, Y., Busta, M., Chowdhury, P. N., Karatzas, D., Khelif, W., Matas, J., Pal, U., Burie, J.-C., Liu, C., & Ogier, J.-M. (2019). ICDAR2019 robust reading challenge on multi-lingual scene text detection and recognition -- RRC-MLT-2019. *arXiv*, (Cornell University).
- Park, S., & Paik, J. (2023). Refcap: image captioning with referent objects attributes. *Scientific Reports*, 13(1), 21577. <https://doi.org/10.1038/s41598-023-48916-6>
- Pustu-Iren, K., Bars, J., Mühling, M., Korfhage, N., Hörth, A., Freisleben, B., & Ewerth, R. (2020). Videomining in historischem material – ein praxisbericht. *Bibliothek Forschung Und Praxis*, 44(3), 436–444. <https://doi.org/10.1515/bfp-2020-2034>
- Rao, Y., Zhao, W., Chen, G., Tang, Y., Zhu, Z., Huang, G., Zhou, J., & Lu, J. (2022). Denseclip: language-guided dense prediction with context-aware prompting. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18061–18070. <https://doi.org/10.1109/CVPR52688.2022.01755>
- Rodrigues da Silva, J., & Caseli, H. de M. (2021). Sense representations for Portuguese: experiments with sense embeddings and deep neural language models. *Language Resources and Evaluation*, 55(4), 901–924. <https://doi.org/10.1007/s10579-020-09525-1>
- Sagduyu, Y. E., Grushin, A., & Shi, Y. (2018). Synthetic social media data generation. *IEEE Transactions on Computational Social Systems*, 5(3), 605–620. <https://doi.org/10.1109/TCSS.2018.2854668>
- Sassu, A., Saenz-Cogollo, J. F., & Agelli, M. (2021). Deep-Framework: a distributed, scalable, and edge-oriented framework for real-time analysis of video streams. *Sensors*, 21(12), 4045. <https://doi.org/10.3390/s21124045>
- Sharma, N., Mandal, R., Sharma, R., Pal, U., & Blumenstein, M. (2015). ICDAR2015 competition on video script identification (CVSI 2015). *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pp. 1196–1200. <https://doi.org/10.1109/ICDAR.2015.7333950>
- Shocher, A., Gandelsman, Y., Mosseri, I., Yarom, M., Irani, M., Freeman, W. T., & Dekel, T. (2020). Semantic pyramid for image generation. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7455–7464. <https://doi.org/10.1109/CVPR42600.2020.00748>

- Souibgui, M. A., Fornés, A., Kessentini, Y., & Megyesi, B. (2022). Few shots are all you need: a progressive learning approach for low resource handwritten text recognition. *Pattern Recognition Letters*, 160, 43–49. <https://doi.org/10.1016/j.patrec.2022.06.003>
- Sun, C., Myers, A., Vondrick, C., Murphy, K., & Schmid, C. (2019). Videobert: a joint model for video and language representation learning. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7463–7472. <https://doi.org/10.1109/ICCV.2019.00756>
- Tao, Y., Jia, Z., Ma, R., & Xu, S. (2021). Trig: transformer-based text recognizer with initial embedding guidance. *Electronics*, 10(22), 2780. <https://doi.org/10.3390/electronics10222780>
- Taurusta, C., Fitriani, A. S., Suryawinata, M., Dijaya, R., & Astutik, M. W. D. (2019). Game education for child with disabled handle based on multimedia. *Journal of Physics: Conference Series*, 1381(1), 012038. <https://doi.org/10.1088/1742-6596/1381/1/012038>
- Tüske, Z., Golik, P., Nolden, D., Schlüter, R., & Ney, H. (2014). Data augmentation, feature combination, and multilingual neural networks to improve ASR and KWS performance for low-resource languages. *Interspeech*, 2014, 1420–1424. <https://doi.org/10.21437/Interspeech.2014-348>
- Vandenhende, S., Georgoulis, S., Van Gansbeke, W., Proesmans, M., Dai, D., & Van Gool, L. (2021). Multi-Task learning for dense prediction tasks: a survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–1. <https://doi.org/10.1109/TPAMI.2021.3054719>
- Veit, A., Matera, T., Neumann, L., Matas, J., & Belongie, S. (2016). COCO-Text: dataset and benchmark for text detection and recognition in natural images. **arXiv preprint arXiv:1601.07140**.
- Wang, L., Zang, J., Zhang, Q., Niu, Z., Hua, G., & Zheng, N. (2018). Action recognition by an attention-aware temporal weighted convolutional neural network. *Sensors*, 18(7), 1979. <https://doi.org/10.3390/s18071979>
- Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., & Loy, C. C. (2019). Esrgan: enhanced super-resolution generative adversarial networks. <https://doi.org/10.1007/978-3-030-11021-5-5>
- Wang, Y., Wang, L., & Su, F. (2018). A robust approach for scene text detection and tracking in video (pp. 303–314). https://doi.org/10.1007/978-3-030-00764-5_28
- Wu, X., Xia, Y., Zhu, J., Wu, L., Xie, S., Fan, Y., & Qin, T. (2021). Mixseq: a simple data augmentation method for neural machine translation. *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, pp. 192–197. <https://doi.org/10.18653/v1/2021.iwslt-1.23>
- Xiang, Y., & Luo, F. (2020). Multi-type web image text detection based on the improved EAST algorithm. *Journal of Physics: Conference Series*, 1544(1), 012115. <https://doi.org/10.1088/1742-6596/1544/1/012115>
- Xu, J., Wei, H., Li, L., Fu, Q., & Guo, J. (2020). Video description model based on temporal-spatial and channel multi-attention mechanisms. *Applied Sciences*, 10(12), 4312. <https://doi.org/10.3390/app10124312>
- Xue, M., Shivakumara, P., Zhang, C., Xiao, Y., Lu, T., Pal, U., Lopresti, D., & Yang, Z. (2021). Arbitrarily-Oriented text detection in low light natural scene images. *IEEE Transactions on Multimedia*, 23, 2706–2720. <https://doi.org/10.1109/TMM.2020.3015037>
- Yan, X., Fang, Z., & Jin, Y. (2023). Augmented Transformers with adaptive n-grams embedding for multilingual scene text recognition. *arXiv*, (Cornell University).
- Yang, A., Miech, A., Sivic, J., Laptev, I., & Schmid, C. (2022). Tubedetr: spatio-temporal video grounding with Transformers. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16421–16432. <https://doi.org/10.1109/CVPR52688.2022.01595>
- Yao, C., Bai, X., Liu, W., Ma, Y., & Tu, Z. (2012). Detecting texts of arbitrary orientations in natural images. *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1083–1090. <https://doi.org/10.1109/CVPR.2012.6247787>
- Yin, W., Wang, H., Qu, J., & Xiong, C. (2021). Batchmixup: improving training by interpolating hidden states of the entire mini-batch. *Findings of the Association for*

- Computational Linguistics: ACL-IJCNLP 2021*, pp. 4908–4912.
<https://doi.org/10.18653/v1/2021.findings-acl.434>
- Yu, H., Zhang, C., Li, X., Han, J., Ding, E., & Wang, L. (2019). An end-to-end video text detector with online tracking. *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pp. 601–606.
<https://doi.org/10.1109/ICDAR.2019.00102>
- Yuan, T.-L., Zhu, Z., Xu, K., Li, C.-J., Mu, T.-J., & Hu, S.-M. (2019). A large Chinese text dataset in the wild. *Journal of Computer Science and Technology*, 34(3), 509–521.
<https://doi.org/10.1007/s11390-019-1923-y>
- Yue, Z., Zeng, H., Lan, M., Ji, H., & Wang, D. (2023). Zero- and few-shot event detection via prompt-based meta learning. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7928–7943.
<https://doi.org/10.18653/v1/2023.acl-long.440>
- Zellers, R., Lu, J., Lu, X., Yu, Y., Zhao, Y., Salehi, M., Kusupati, A., Hessel, J., Farhadi, A., & Choi, Y. (2022). Merlot RESERVE: neural script knowledge through vision and language and sound. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16354–16366.
<https://doi.org/10.1109/CVPR52688.2022.01589>
- Zhan, F., & Lu, S. (2019). Esir: end-to-end scene text recognition via iterative image rectification. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2054–2063.
<https://doi.org/10.1109/CVPR.2019.00216>
- Zhou, Q., Li, X., He, L., Yang, Y., Cheng, G., Tong, Y., Ma, L., & Tao, D. (2023). Transvod: end-to-end video object detection with spatial-temporal Transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6), 7853–7869.
<https://doi.org/10.1109/TPAMI.2022.3223955>
- Zhou, X., Yao, C., Wen, H., Wang, Y., Zhou, S., He, W., & Liang, J. (2017). East: an efficient and accurate scene text detector. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2642–2651.
<https://doi.org/10.1109/CVPR.2017.283>
- Ziat, A., Delasalles, E., Denoyer, L., & Gallinari, P. (2017). Spatio-Temporal neural networks for space-time series forecasting and relations discovery. *2017 IEEE International Conference on Data Mining (ICDM)*, pp. 705–714.
<https://doi.org/10.1109/ICDM.2017.80>