



A Comprehensive Review On Text Signage Recognition And Translation System For Indian Languages

¹Manjula B.K,²Latha B.M.

¹Assistant Professor,²Professor,

¹Dept. of Electronics and Communication Engineering,²Dept. of Information Science & Engineering.

¹G.M. Institute of Technology, Davangere, India. ²Jain Institute of Technology, Davangere, India.

Abstract: Scene text recognition on outdoor signboards, particularly in Indian languages, is a crucial application of deep learning with broad implications for navigation, automated translation, and accessibility. Despite advancements, challenges such as lighting variations, occlusions, distortions, and complex backgrounds hinder accurate text extraction. This study presents a deep learning-based framework for text signage recognition and translation, specifically tailored for Indian languages. The proposed methodology involves a multi-stage approach, including data preprocessing for noise reduction and contrast enhancement, feature extraction using a Feature Pyramid Network, and text recognition through a Bidirectional Long Short-Term Memory network. A hybrid Convolutional Neural Network - Recurrent Neural Network model is employed to classify and convert extracted text into a machine-readable format, followed by translation using transformer-based models trained on parallel Indian language corpora. This system aims to enhance multilingual text recognition and real-time translation, contributing to improved accessibility and navigation for users across diverse linguistic backgrounds.

Index Terms - Scene text recognition, Deep learning, Feature Pyramid Network, Bidirectional LSTM, Convolutional Neural Network, Recurrent Neural Network, Text extraction, Multilingual translation, Indian languages.

I. INTRODUCTION

Scene text detection is a fundamental step in text recognition and plays a vital role in various applications, including image retrieval, machine translation, and autonomous driving. With the widespread adoption of internet technology and the increasing use of mobile devices, extracting text from images and videos has become an essential task. Over the past decade, numerous methods have been developed to detect and recognize text in natural scenes, yielding promising results. As a result, scene text detection and recognition have emerged as key areas of research in computer vision and pattern recognition, distinct from traditional document analysis [1].

Despite notable progress, several challenges persist. One major difficulty is the diversity and variability of text in natural scenes. Unlike structured documents, scene text appears in multiple languages and scales, with variations in shape, orientation, and color. These differences significantly complicate the detection process. Additionally, background elements such as signboards, fences, bricks, or grass often share visual similarities with text, making it difficult to distinguish them accurately. Another challenge is occlusion, where parts of the text are obstructed by external objects, leading to potential recognition errors. Furthermore, image quality can be inconsistent due to varying capture conditions. Factors such as distortion, out-of-focus shots, and lighting variations can introduce noise, shadows, and blurriness, all of which impact detection accuracy. Analyzing text in natural scene images presents unique difficulties, particularly in recognizing words on billboards, road signs, and other environmental objects. Successfully extracting such text enables numerous applications, including content-based image retrieval, sign translation, intelligent

driving assistance, and navigation support for visually impaired individuals and robotic systems. Due to these diverse applications, scene text detection has gained significant research interest in recent years [2].

Scene Text Detection and Recognition (STDR) refers to identifying and localizing textual content within images captured in real-world environments. Beyond mere detection, STDR plays a key role in extracting high-level semantic information, making it valuable in fields such as risk and knowledge management, cybersecurity, content enhancement, and fraud prevention. Additionally, it facilitates information retrieval from images and videos at different levels, from entire pages to individual words and characters. Once text is detected, recognition becomes essential for various computer vision applications, including automatic sign reading, self-driving vehicle navigation, real-time language translation, and multimedia content retrieval. A notable example of its application is in driver-assistance systems and autonomous vehicles, where this technology significantly enhances passenger safety and overall security [3].

STDR datasets are often divided into two categories: regular and irregular. Regular datasets contain images where text is aligned horizontally, making detection and recognition relatively straightforward. In contrast, irregular datasets include text that is distorted, curved, or arbitrarily rotated, posing additional challenges. Furthermore, scene text can be categorized based on its form as either handwritten or printed. Another classification is based on language, where text is grouped into Latin and non-Latin scripts. In many cases, images may contain multiple languages within a single frame as shown in Figure 1, further increasing the complexity of recognition tasks.

The presence of such variations in real-world STDR data requires robust techniques capable of handling irregular text orientations, diverse scripts, and complex backgrounds. These challenges highlight the need for advanced models that can efficiently process and recognize text in diverse conditions, ensuring accurate and reliable STDR performance.



Figure 1. Sample image sources for scene text recognition in Indian languages, showcasing multilingual text in real-world scenarios.

Traditional text detection approaches include texture-based methods [4-6], which treat text as a distinct texture, and region-based methods [7-11], which extract candidate components. However, these techniques are often sensitive to changes in rotation and scale. Recently, deep learning-based approaches [12-18] have gained traction, demonstrating significant improvements. By training deep neural networks to

automatically extract features rather than relying on manually designed extractors, these methods have enhanced text detection accuracy.

To address these challenges, we propose a text proposal network that leverages a feature pyramid structure, integrating Convolutional Neural Networks [19] and Recurrent Neural Networks for text localization. Our approach combines multi-resolution features from different convolutional layers into a feature pyramid, ensuring a robust representation of text at various scales. Additionally, a bidirectional Long Short-Term Memory (Bi-LSTM) network is employed to encode these features, capturing contextual information and reducing errors associated with detecting individual characters. Finally, a specialized text connector is designed to link detected text proposals into complete text lines, regardless of orientation.

II LITERATURE REVIEW

Over the past decade, numerous methods have been developed for text detection in natural scenes. These methods generally fall into two main categories: traditional bottom-up approaches and deep learning-based top-down techniques. Bottom-up methods typically follow a sequential process, identifying individual character components before grouping them into words or text regions. In contrast, top-down approaches leverage deep learning models to directly detect text in images, often achieving higher accuracy and robustness.

Texture-based methods [20]–[22] consider text as a distinct form of texture for analysis. Some approaches [20] utilize the Discrete Cosine Transform to examine the directivity and periodicity of an image block, aiding in text localization. Others [21] employ the Fourier spectrum to determine the fundamental frequency of a text image, leveraging frequency characteristics for text area identification. Additionally, Gabor filters have been applied to extract four stroke features of Chinese characters, reframing the text detection task as a texture classification problem, which is then processed using a Support Vector Machine [22].

Region-based approaches [23]–[25] identify candidate components and eliminate non-text elements using filters or classifiers. By focusing on candidate regions, these methods reduce the processing area compared to texture-based techniques, making text features more prominent within smaller regions. This enhances the contrast between text and non-text areas, significantly improving detection efficiency. Among the widely used methods, Stroke Width Transform (SWT) [23] computes the stroke width for each pixel and extracts text candidate regions across various scales and orientations, even in complex backgrounds. Huang et al. [25] introduced Stroke Feature Transformation (SFT), an extension of SWT, incorporating an SFT filter to segment and link components effectively. Additionally, Neumann and Matas [24] employed Maximally Stable Extremal Regions to detect candidate text regions, followed by a classifier to refine the selection.

Hybrid approaches [26], [27] integrate both texture-based and region-based methods, leveraging the strengths of each. Liu et al. [26] introduced a technique that combines connected component analysis with texture feature extraction to analyze contours of unidentified text regions. Pan et al. [27] developed a text region detector capable of estimating text confidence and scale information within an image pyramid. This facilitates the segmentation of candidate text components through local binarization, while conditional random fields are used to classify components as text or non-text.

Despite their advantages, texture-based approaches [20]–[22] have limitations, including simplistic feature construction and low detection accuracy. Enhancing accuracy often requires numerous predefined rules and parameter constraints, leading to reduced generalization and robustness. Similarly, while region-based methods [23]–[25] can identify text candidate regions across various scales and orientations in complex backgrounds, they still depend on manually crafted rules for effective text detection. Even hybrid approaches [26], [27], which merge the strengths of both techniques, struggle with adaptability to multi-oriented text and variations in scale.

With the rapid advancement of deep learning, its applications in computer vision have become increasingly prominent. Currently, top-down approaches [28]–[34] based on Convolutional Neural Networks are among the most widely used methods. Deep learning techniques have significantly enhanced text detection accuracy while eliminating the need for complex feature engineering. Through multiple convolutional layers, these models extract abstract and high-level features, enabling more effective classification of text and background. Additionally, neural networks utilize bounding box regression to accurately localize text regions, further improving detection performance.

In recent studies, state-of-the-art object detection methods such as Faster R-CNN [37], Single Shot MultiBox Detector [38], and Fully Convolutional Networks (FCN) [39] have gained significant attention

from researchers. Inspired by these techniques, Ma et al. [29] proposed Rotation Region Proposal Networks (RRPN), which incorporated rotation factors into Faster R-CNN. RRPN generates inclined proposals with text orientation angle information, replacing the Region-of-Interest (RoI) pooling layer of Faster R-CNN with a Rotation Region-of-Interest (RRoI) pooling layer to handle arbitrary-oriented proposals. Liao et al. [30] introduced an end-to-end trainable fast scene text detector called TextBoxes, based on SSD, which eliminates post-processing except for standard non-maximum suppression, making it highly efficient. To address the limitation of TextBoxes in detecting tilted text, they developed TextBoxes++ [31], which improved detection by incorporating rotated rectangles or quadrilaterals.

Liu et al. [24] proposed a new framework for arbitrarily oriented scene text detection using FCN, where two FCNs were used to detect text center blocks and word stroke regions, followed by a word region surrounding box algorithm for final detection. The Connectionist Text Proposal Network [34] is a well-known approach for detecting horizontal text. It identifies text lines by generating a sequence of fine-scale text proposals directly from convolutional feature maps. Liu and Jin [28] introduced the Deep Matching Prior Network, which enhances text detection using a tighter quadrangle representation based on CNNs. This method employs a quadrilateral sliding window to improve recall and utilizes a shared Monte Carlo technique for calculating polygonal areas.

There are also many methods that detect and recognize text using end-to-end trainable models. The Robust Text Recognizer with Automatic Rectification (RARE) [32] consists of a Spatial Transformer Network and a Sequence Recognition Network, enabling effective recognition of perspective and curved text. Additionally, Shi et al. [33] proposed a Convolutional Recurrent Neural Network, which can handle sequences of arbitrary lengths by integrating feature extraction, sequence modeling, and transcription into a unified framework.

While previous methods have made significant advancements in detecting multi-scale and multi-oriented text lines, challenges related to efficiency and precision remain. Our approach leverages Feature Pyramid Networks [40], which effectively addresses multi-scale text detection without adding computational complexity. Additionally, when handling multi-oriented text, our method reduces the number of required parameters per anchor compared to traditional schemes. Unlike the five-parameter rotation detection box and the eight-parameter quadrilateral detection box, our approach requires only two parameters, significantly optimizing the training process and improving overall performance.

Table 1 presents a comparative analysis of various text detection methods, highlighting their descriptions, advantages, and limitations. This comparative analysis aids in understanding the trade-offs between different text detection techniques, enabling us to choose the most suitable approach for the specific applications in text signage recognition and translation systems.

Table 1. Comparison of Text Detection Methods- Description, Advantages and Limitations

Sl. No	Paper	Method	Description	Advantages	Limitations
1.	Y.Zhong et al. [20]	Discrete Cosine Transform	Examines directivity and periodicity of image blocks for text localization	Effective for structured text	Simplistic feature construction, low accuracy
2.	B.K.Sin et al. [21]	Fourier Spectrum	Determines fundamental frequency of text images for text area identification	Utilizes frequency characteristics effectively	Requires parameter tuning, low robustness
3.	J. Yan et al. [22]	Gabor Filters + SVM	Extracts stroke features of Chinese characters for texture classification	Works well with structured text	Requires handcrafted features, lacks adaptability
4.	B. Epshtein et al. [23]	Stroke Width Transform	Computes stroke width for each pixel to extract text regions	Works in complex backgrounds and multiple orientations	Relies on predefined rules

5.	L. Neumann et al. [24]	Maximally Stable Extremal Regions	Detects text regions using stable intensity regions	Effective for high contrast text	Sensitive to noise and illumination changes
6.	W. Huang et al. [25]	Stroke Feature Transformation	Enhances SWT with filtering for better segmentation	Improves text segmentation accuracy	Still depends on handcrafted rules
7.	Y. Liu et al. [26]	Connected Component Analysis + Texture Features	Combines region-based and texture-based approaches	Improves text localization	Limited adaptation to scale variations
8.	Y.F. Pan et al. [27]	Conditional Random Fields	Uses local binarization and CRF to classify text	Better segmentation of text components	Struggles with multi-oriented text
9.	Y. Liu et al. [28]	Deep Matching Prior Network	Uses quadrilateral sliding windows for text localization	Improves recall	Complex computation
10.	J. Ma et al. [29]	Rotation Region Proposal Networks	Integrates rotation factors into Faster R-CNN for inclined text	Detects multi-oriented text effectively	High computational cost
11.	M. Liao et al. [30]	TextBoxes	End-to-end SSD-based scene text detector	Eliminates post-processing	Struggles with tilted text
12.	M. Liao et al. [31]	TextBoxes++	Improves TextBoxes with rotated rectangles and quadrilaterals	Enhances tilted text detection	Increased model complexity
13.	B. Shi et al. [32]	Robust Text Recognizer with Automatic Rectification	Uses Spatial Transformer Network and Sequence Recognition Network	Handles curved and perspective text	Requires high computational power
14.	B. Shi et al. [33]	Convolutional Recurrent Neural Network	Integrates feature extraction, sequence modeling, and transcription	Handles arbitrary text sequences	Dependent on training data quality
15.	Z. Tian et al. [34]	Connectionist Text Proposal Network	Generates fine-scale text proposals for horizontal text	Accurate text line detection	Limited to horizontal text
16.	S. Ren et al. [37]	Faster R-CNN	Deep learning-based object detection for text localization	High accuracy, no feature engineering needed	Computationally expensive
17.	W. Liu et al. [38]	Single Shot MultiBox Detector	Uses deep learning to detect text in a single shot	Fast and efficient	Less effective for small text
18.	J. Long et al. [39]	Fully Convolutional Networks	Uses FCN to detect text center blocks and stroke regions	Handles arbitrary-oriented text	Complex training process
19.	T.Y. Lin et al. [40]	Feature Pyramid Networks	Uses multi-scale feature maps for text detection	Efficient multi-scale detection	Parameter optimization required

III PROPOSED METHODOLOGY

Figure 2. illustrates the proposed methodology for developing a text signage recognition and translation system tailored for Indian languages, leveraging deep learning techniques. A Text Signage Recognition and Translation System for Indian Languages using deep learning involves multiple stages: data acquisition, feature extraction, text recognition, and translation. First, a diverse dataset of signboard images in Indian languages like Kannada, Hindi, Tamil, and Telugu is collected and preprocessed using techniques such as noise removal, contrast adjustment, and data augmentation to enhance recognition accuracy. A **Feature Pyramid Network** extracts multi-scale features, ensuring effective text detection from complex backgrounds. The extracted text regions are then processed using **Bidirectional Long Short-Term Memory** to capture contextual dependencies in both forward and backward directions, improving recognition accuracy.

Region of Interest (ROI) Pooling isolates and normalizes text regions for better segmentation, enhancing model efficiency. A deep learning-based optimized **CNN-RNN hybrid model** then classifies and converts text into a machine-readable format. The recognized text is translated into a target language using transformer-based models, trained on parallel Indian language corpora. Finally, the system's performance is evaluated using **Character Error Rate, Word Error Rate, BLEU, and METEOR** scores. This system aims to break language barriers, offering real-time signboard translation for tourists and non-native speakers, enhancing accessibility and navigation.

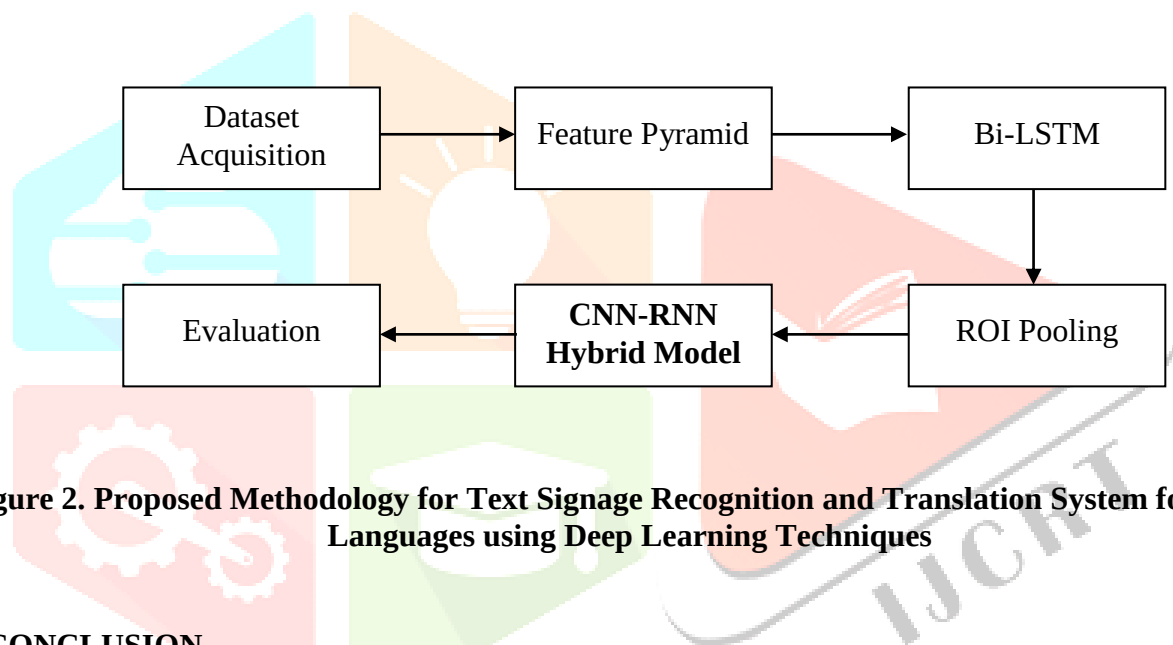


Figure 2. Proposed Methodology for Text Signage Recognition and Translation System for Indian Languages using Deep Learning Techniques

IV. CONCLUSION

Scene text recognition, particularly for Indian languages, faces challenges due to variations in text orientation, script complexity, and environmental factors like lighting and occlusion. This review explored traditional and deep learning-based approaches, highlighting their strengths and limitations. Traditional methods, including texture-based and region-based techniques, perform well in structured settings but struggle with real-world complexities. Deep learning models, especially CNNs and RNNs, have significantly improved recognition accuracy by learning robust feature representations. The proposed deep learning framework enhances recognition through advanced preprocessing, segmentation, and optimized models for multilingual scenarios. With growing applications in autonomous driving, accessibility for visually impaired individuals, and real-time translation, improving model efficiency is crucial. Future research should focus on handling low-resource languages and optimizing recognition under extreme conditions. Addressing these challenges will contribute to more reliable and efficient scene text recognition systems. Such advancements will enhance accessibility and information retrieval across diverse linguistic landscapes. The continued evolution of AI-driven text recognition will further expand its real-world applications.

REFERENCES

- [1] Wang, Q., Huang, Y., Jia, W., He, X., Blumenstein, M., Lyu, S., & Lu, Y. (2020). FACLSTM: ConvLSTM with focused attention for scene text recognition. *Science China Information Sciences*, 63, 1-14.
- [2] Lyu, P., Yao, C., Wu, W., Yan, S., & Bai, X. (2018). Multi-oriented scene text detection via corner localization and region segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7553-7563.
- [3] Liu, F., Chen, C., Gu, D., & Zheng, J. (2019). FTPN: scene text detection with feature pyramid based text proposal network. *IEEE Access*, 7, 44219-44228. <https://ieeexplore.ieee.org/document/8682040>.
- [4] Y. Zhong, H. Zhang, and A. K. Jain, "Automatic caption localization in compressed video," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 4, pp. 385–392, Apr. 2000. doi: 10.1109/34.845381.
- [5] B.-K. Sin, S.-K. Kim, and B.-J. Cho, "Locating characters in scene images using frequency features," in *Proc. Object Recognit. Supported Interact. Service Robots*, Quebec City, QC, Canada, Aug. 2002, pp. 489–492.
- [6] J. Yan, J. Li, and X. Gao, "Chinese text location under complex back-ground using Gabor filter and SVM," *Neurocomputing*, vol. 74, no. 17, pp. 2998–3008, 2011. doi: 10.1016/j.neucom.2011.04.031.
- [7] B. Epshtein, E. Ofek, and Y. Wexler, "Detecting text in natural scenes with stroke width transform," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, San Francisco, CA, USA, Jun. 2010, pp. 2963–2970.
- [8] L. Neumann and J. Matas, "A method for text localization and recognition in real-world images," in *Proc. Asian Conf. Comput. Vis.* Berlin, Germany: Springer, 2010, pp. 770–783.
- [9] W. Huang, Z. Lin, J. Yang, and J. Wang, "Text localization in natural images using stroke feature transform and text covariance descriptors," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1241–1248.
- [10] Y. Liu, S. Goto, and T. Ikenaga, "A contour-based robust algorithm for text detection in color images," *IEICE Trans. Inf. Syst.*, vol. E89-D, no. 3, pp. 1221–1230, 2006. doi: 10.1093/ietisy/e89-d.3.1221.
- [11] Y.-F. Pan, X. Hou, and C.-L. Liu, "A hybrid approach to detect and localize texts in natural scene images," *IEEE Trans. Image Process.*, vol. 20, no. 3, pp. 800–813, Mar. 2011. doi: 10.1109/tip.2010.2070803.
- [12] Y. Liu and L. Jin, "Deep matching prior network: Toward tighter multi-oriented text detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 3454–3461.
- [13] J. Ma et al., "Arbitrary-oriented scene text detection via rotation proposals," *IEEE Trans. Multimedia*, vol. 20, no. 11, pp. 3111–3122, Nov. 2018.
- [14] M. Liao, B. Shi, X. Bai, X. Wang, and W. Liu, "TextBoxes: A fast text detector with a single deep neural network," in *Proc. AAAI*, 2017, pp. 4161–4167.
- [15] M. Liao, B. Shi, and X. Bai, "TextBoxes++: A single-shot oriented scene text detector," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3676–3690, Aug. 2018.
- [16] B. Shi, X. Wang, P. Lyu, C. Yao, and X. Bai, "Robust scene text recognition with automatic rectification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 4168–4176.
- [17] B. Shi, X. Bai, and C. Yao, "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 11, pp. 2298–2304, Nov. 2017.
- [18] Z. Tian, W. Huang, T. He, P. He, and Y. Qiao, "Detecting text in natural image with connectionist text proposal network," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 56–72.
- [19] K. Simonyan and A. Zisserman. (2014). "Very deep convolutional networks for large-scale image recognition." [Online]. Available: <https://arxiv.org/abs/1409.1556>.
- [20] Y. Zhong, H. Zhang, and A. K. Jain, "Automatic caption localization in compressed video," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 4, pp. 385–392, Apr. 2000. doi: 10.1109/34.845381.
- [21] B.-K. Sin, S.-K. Kim, and B.-J. Cho, "Locating characters in scene images using frequency features," in *Proc. Object Recognit. Supported Interact. Service Robots*, Quebec City, QC, Canada, Aug. 2002, pp. 489–492.
- [22] J. Yan, J. Li, and X. Gao, "Chinese text location under complex background using Gabor filter and SVM," *Neurocomputing*, vol. 74, no. 17, pp. 2998–3008, 2011. doi: 10.1016/j.neucom.2011.04.031.

- [23] B. Epshtein, E. Ofek, and Y. Wexler, "Detecting text in natural scenes with stroke width transform," in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., San Francisco, CA, USA, Jun. 2010, pp. 2963–2970.
- [24] L. Neumann and J. Matas, "A method for text localization and recognition in real-world images," in Proc. Asian Conf. Comput. Vis. Berlin, Germany: Springer, 2010, pp. 770–783.
- [25] W. Huang, Z. Lin, J. Yang, and J. Wang, "Text localization in natural images using stroke feature transform and text covariance descriptors," in Proc. IEEE Int. Conf. Comput. Vis., Dec. 2013, pp. 1241–1248.
- [26] Y. Liu, S. Goto, and T. Ikenaga, "A contour-based robust algorithm for text detection in color images," IEICE Trans. Inf. Syst., vol. E89-D, no. 3, pp. 1221–1230, 2006. doi: 10.1093/ietisy/e89-d.3.1221.
- [27] Y.-F. Pan, X. Hou, and C.-L. Liu, "A hybrid approach to detect and localize texts in natural scene images," IEEE Trans. Image Process., vol. 20, no. 3, pp. 800–813, Mar. 2011. doi: 10.1109/tip.2010.2070803.
- [28] Y. Liu and L. Jin, "Deep matching prior network: Toward tighter multioriented text detection," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Honolulu, HI, USA, Jul. 2017, pp. 3454–3461.
- [29] J. Ma et al., "Arbitrary-oriented scene text detection via rotation proposals," IEEE Trans. Multimedia, vol. 20, no. 11, pp. 3111–3122, Nov. 2018.
- [30] M. Liao, B. Shi, X. Bai, X. Wang, and W. Liu, "TextBoxes: A fast text detector with a single deep neural network," in Proc. AAAI, 2017, pp. 4161–4167.
- [31] M. Liao, B. Shi, and X. Bai, "TextBoxes++: A single-shot oriented scene text detector," IEEE Trans. Image Process., vol. 27, no. 8, pp. 3676–3690, Aug. 2018.
- [32] B. Shi, X. Wang, P. Lyu, C. Yao, and X. Bai, "Robust scene text recognition with automatic rectification," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Las Vegas, NV, USA, Jun. 2016, pp. 4168–4176.
- [33] B. Shi, X. Bai, and C. Yao, "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition," IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 11, pp. 2298–2304, Nov. 2017.
- [34] Z. Tian, W. Huang, T. He, P. He, and Y. Qiao, "Detecting text in natural image with connectionist text proposal network," in Proc. Eur. Conf. Comput. Vis. Cham, Switzerland: Springer, 2016, pp. 56–72.
- [35] K. Simonyan and A. Zisserman. (2014). "Very deep convolutional networks for large-scale image recognition." [Online]. Available: <https://arxiv.org/abs/1409.1556>
- [36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2016, pp. 770–778.
- [37] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards realtime object detection with region proposal networks," IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 6, pp. 1137–1149, Jun. 2017. doi: 10.1109/tpami.2016.2577031.
- [38] W. Liu et al., "SSD: Single shot multibox detector," in Proc. Eur. Conf. Comput. Vis., 2016, pp. 21–37.
- [39] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 3431–3440.
- [40] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 936–944.