Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.777226

TEXT LOCALIZATION AND RECOGNITION IN VIDEO AND IMAGES

Suman
Research scholar, Department of Computer Science and Engineering, University Visvesvaraya College of Engineering, Bangalore University, Bengaluru-560001
Sumanjosh92@gmail.com

Dr. H N Champa
Professor, Department of Computer Science and Engineering, University Visvesvaraya College of Engineering, Bangalore University, Bengaluru-560001
champahn@uvce.ac.in

Abstract—Computer vision's many applications include sporting related audio-visual analysis, autonomous driving, and industrial robotics to mention a few, all needing the localization and recognition of text in natural images. They have to deal with a wide range of issues affecting how text is displayed and influenced by the surrounding environment. Since deep learning architectures have seen significant advancements in recent years, contemporary scene text detection and identification approaches report higher accuracy on benchmark datasets when dealing with multi-resolution, multi-oriented text. On the other hand, existing systems can still not generalize to unseen and insufficiently labeled data, which causes them to underperform in the bleakness photos. Systematic survey is carried out on methods applied for localizing the text in video and images.
Keywords—Artificial Intelligence, Image Localization, Natural language Processing

I. INTRODUCTION

Text is a set of cyphers that can be used to preserve, transmit, or propagate culture. Text, one of humanity's most important inventions, has played a significant role in our lives. In a diverse variety of perception-based application scenarios, correct semantic information conveyed by text is critical. When Multimedia elements are enhancing information more and more, huge digital picture and video libraries are also developing, and libraries that were formerly exclusively text-based are constantly adding photos, movies, and audio samples to their repositories. They all require an automated method to effectively index and retrieve multimedia components. One potential source of high-level semantics is text in pictures, particularly text that appears in images that are part of online sites and movies. These text occurrences would be a valuable source of high-level concepts for indexing and retrieval if they could be automatically found, split, and recognized. For instance, movie text occurrences are a crucial information source for the Informedia Project, enabling a thorough exploration and innovation of their terabyte digital video archive of broadcasts and life stories [1]. Identification and understanding of common text patterns could also be used to record the date and time of advertisement assisting individuals in determining if their client's advertising have indeed been aired at the scheduled time on the scheduled television channel [2].

Not only on television and other media, the text is also presented in images on more and more web pages, making it crucial to identify, detect, and segment text in these images. The text cannot be extracted using text-segmentation and text-recognition methods currently in use. As a result, none of the search engines in use can accurately index the content of websites with plenty of images [3]. Since the textual content in images may be extracted, automatic text segmentation and text recognition also aid in the automatic translation of web pages made for big monitors to small LCD appliances.

Given the increased importance and widespread use of text in different media platforms, much research has been conducted in the field of text recognition on videos and images. At the character (analytical) and holistic levels of character recognition in document pictures, a variety of strategies have been proposed. This allows for the text to be recognized. Graph-based models [4, 5], Bayesian classifiers [6, 7], and hidden Markov models (HMMs) [8, 9], among others, have frequently been examined. These are all examples of systems that use characters as recognition units. Studies [10-12] on a broad array of characteristics and classifiers have reported good identification rates when used in conjunction with holistic or word-level recognition approaches. In addition, deep learning has been used for characteristics extraction and categorization at the level of individual characters and words [13].

One recent research conducted on this issue includes that of Zayeneet al. (2019). They developed a segmentation-free strategy based on multi-dimensionallong short-term memory(MDLSTM) for the recognition of Arabic video text [14]. The suggested method achieves high identification rates on two data sets that are used as benchmarks. These data sets are ACTiV and ALIF [16]. Another investigation was conducted by Jain et al. (2022), employing CNN–LSTM hybrid DNN. They used the identical data sets as the previous study [15]. Here, a noteworthy technique for audio-visual text recognition is provided. In this method, an SVM is used to pick an appropriate color channel, and then HMMs are used to recognize text based on the selected color channel. For video text recognition, Lu et al. (2019) use transfer learning in conjunction with pre-trained CNNs [16]. In this particular investigation, cutting-edge models such as InceptionV3, VGG16, and Resnet50 are taken into consideration. Xu et al. (2018) present a similar study for recognizing East Asian languages. This study uses CNNs to recognize Chinese characters [17].

Kwang In Kim et al.(2003) this paper aims to provide a detailed study of text detection, tracking, and identification systems in a single frame (images) and video, with a particular emphasis on recent advances in technological capabilities.

II. METHODS

Text detection approaches can be classified into two parts. One is the classic machine learning techniques, and the other is the modern deep learning approaches. In Fig. 1, various kinds of classic ML and deep learning procedures are presented in a flow diagram.

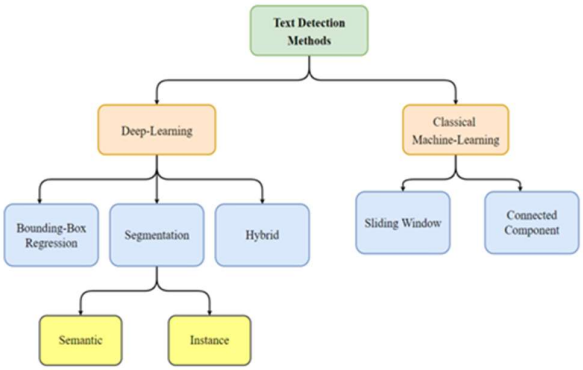

Fig.1 Flow diagram of text detection methods

Machine learning techniques:

It is possible to classify these techniques in two primary ways: the connected-component-based approach and the sliding-window approach.

An image pyramid is built from a specified test image and then scanned over all potential text places and scales through a sliding window of a specific size in approaches such as [18-20]. Once this is done, a classical classifier is used to classify each window using a variety of weak classifiers, such as decision trees, log-likelihood, likelihood ratio t, and histograms of oriented gradients (HOG), as well as the histograms to extract a specific type of image feature from each window. For example, Chen et al. [20] analyzed intensity histograms, gradients, and its directions at each sliding window location in the test image. AdaBoost, a framework for text detection, was used to combine numerous weak log-likelihood classifiers trained on the same sort of data into a more robust classifier. The non-maximal suppression (NMS) was implemented to determine each character separately. HOG features were retrieved at each sliding window region, and a Randomized Fern classifier was used for multi-scale character identification. Although these algorithms can detect only horizontal text, they have a poor detection accuracy in on-scene photos containing text in any orientation.

The connected component main aim to identify the regions with matched features and features can be color, texture and boundary. to generate selection elements that can be characterized into text or non-text classes that use a conventional classification model (such as support vector machine [21], Random Forest [22], and nearest-neighbor[23]). There are many ways to extract characters from a picture and afterward merge them into words or lines of text. For scene text identification, the reduced false-positive rate of connected-component approaches is critical since they outperform sliding-window methods in terms of efficiency and robustness [24].

Modern techniques based on Deep Learning Systems:

As deep learning has grown in popularity, it has altered the way academics approach text detection problems and greatly broadened the field's potential research applications [25]. Using deep learning-based techniques (such as a quicker and modest pipeline, detecting typescript of varied aspect ratios [26], and offering the potential to be educated efficiently using simulated data [27]) has been widely employed.

Text detection techniques based on deep learning that was developed earlier [28],[29],[27],[30] often involve several phases. For example, the CNN system was modified in terms of training a classification algorithm to build a text saliency map. Afterward, bounding boxes were integrated at several sizes by conducting filtering and NMS. Huang et al[28], combined deep learning with the more traditional connected component-based technique. to achieve a higher level of accuracy with their final text detector [28]. In this method, a conventional MSER [31] high contrast regions detector was applied to the input image to search for character candidates. Next, a CNN classifier was applied to the image to filter out non-text representatives by creating a confidence map, which was then utilized for the purpose of obtaining the detection results. Finally, the MSER high contrast regions detector was applied to the participation picture to find character contenders.

Later on in [27], the overall determining feature Auto Correlation Function (ACF) detector [32] was utilized to create text candidates. After that, a CNN was employed for bounding box regression in order to reduce the amount of candidates who are mistakenly identified as positive. Even so, these techniques [28], [29], [30]were primarily used to detect characters; as a result, their performance may suffer when characters are present within a complicated background, such as when elements of the background have an appearance that is comparable to that of the characters, or when characters are affected by geometric variations [33].

Approaches for text detection that are based on bounding-box regression view text as an object and try to directly forecast candidate bounding boxes. These methods are used to detect text. Text Boxes, for instance[34], change the single-shot descriptor (SSD) [35], kernels by employing long default anchors and filters to manage the considerable fluctuation of aspect ratios seen within text instances. This was done so that the kernels could better cope with the situation. In Shi et al. (2017), they used an architecture inherited from SSD to decompose text into reduced parts and after that join them into text instances, a process they call SegLink[26]. SegLink does it using spatial relations or linking predictions between neighbouring text segments, which enabled SegLink to detect huge queues of Latin and non-Latin text having high expansion ratios. The Connectionist Text Proposal Network (CTPN) [36], a revamped form of FasterRCNN[37], utilised an anchor mechanism to simultaneously forecast the place and mark of every fixed-width proposal. Following this, a recurrent neural network was used to connect the sequential proposals (RNN). Gupta et al[38],developed a entirely convolutional regression network [38] that was stimulated by the YOLO network [39]. Additionally, a random-forest classifier was applied in order to cut down on the amount of false-positive text that was detected in images.

III. SINGLE FRAME TEXT RECOGNITION

The comprehensive and accurate information delivered by text is essential in various vision-based application settings. It is difficult, however, to pull text from photographs of natural settings and then use that text in another program. Text localization [40],[41],[42],[43], text verifying [44, 45], [46], text recognition [22], [47], text edge detection [48],[49],[19],[50], optical character recognition [51],[52],[53], and end-to-end systems [54], [55],[46] are some of the major glitches that were demarcated at different legs of this chore in the collected works.

In addition, certain text-related concerns emerge since the challenges posed by language are unlike anything else. Some are text augmentation, text tracking, natural language processing (NLP), etc. Researchers can benefit from a solid understanding of these basic concepts to better assess the variations and interconnections between various activities.

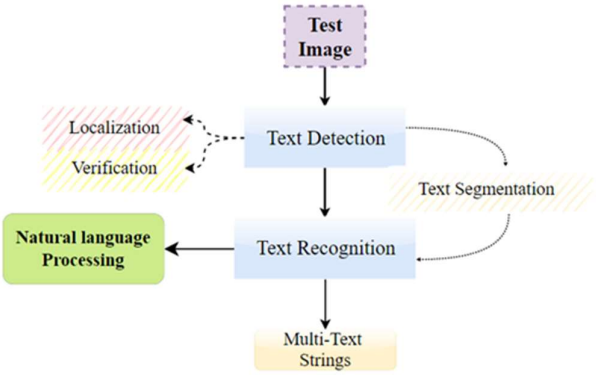

Fig. 2 simple flow diagram for Image -Text- Detection

In Fig. 2 various issues and approaches of text image detection and recognition are represented in a flow diagram.

Text localization:

Image localization is a predictive model in which the outcome is a set of x and y coordinates surrounding the entity that may be used to build bounding boxes.

Text verification:

The text verification process aims to control if a text candidate region contains text or not. Because text localization can occasionally result in false positives, this step is typically performed after text localization in filtering the contender regions.

Text detection:

This is used to regulate whether text is present or not by employing localization and verification courses. It offers accurate and condensed text instance images for the purpose of text recognition, serving as the foundation of an end-to-end structure.

Text segmentation:

Text segmentation is a difficult problem to solve. It comprises separating lines of text and characters. The first means is to divide a region with more than one line of text into smaller regions with only one line of text each. This denotes to severe a block of transcript into manifold areas with lone characters each. Character segmentation was often utilized in initial attempts at text recognition.

Text recognition:

When a text instance image is cropped, it is translated to the target string sequence via text recognition. End-to-end systems that offer reliable outcomes rely on this component.

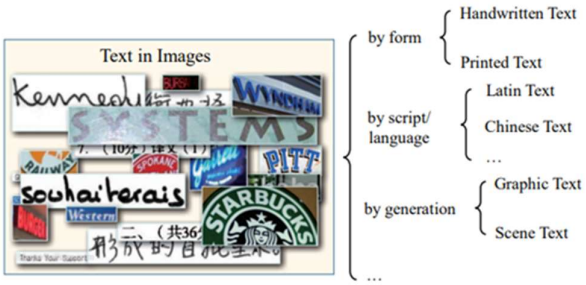

Fig. 3 Categorizations of text in a single frame (images)[56]

Text can appear differently in the images. Figure 3 shows examples and typical classifications. For example, if classified by the text form, handwritten text and printed text are two basic classes. Notably, classification methods may overlap. Handwritten text recognition is more challenging than printed text recognition because of various handwriting styles and character-touching problem.

The End-to-end system:

It is possible to directly transform all text regions in a scene image into the target string sequences using an end-to-end system. This procedure includes all the stages mentioned above.

Natural language processing:

Human languages can be decoded and analysed by computer programmes. This procedure utilises Natural Language Processing technology. NLP serves as an intermediary between humans and computers.Named Entity Recognition(NER), speech recognition, relationship extraction, and topic segmentation are just a few of the activities that may be accomplished with the help of this tool. Its primary focus is text, the most prevalent unstructured data type.

Crucial semantic information about an image can sometimes be found in the text present in said frame or image. The detection of this text is a challenge that researchers over the years have been trying to solve using image processing and deep learning methods. The problems with successful text detection and localization in multimedia frames is that it is often arbitrarily oriented, it may or may not have a complex background that makes the text difficult to decipher, and the fact that it is often not uniform in shape, size, or colour [57]. An effective text localization and recognition model should detect text despite these issues.

Fig. 4 Iterative Nearest Neighbor Symmetry (INNS) and Mutual Nearest Neighbor Pair (MNNP) components

based on gradient direction [58]
One such example is shown in Fig. 4, where from the input image it is clearly shows bit 5 and bit 4 gives complete details compare to bit 7 and bit 6. Author proposed a mechanism to recognise the plan in the subsequent segments. Then extract the bits from byte of pixel data from the gray input image to retrieve the bit plane slices (images), the slices provide significant information in binary form. Therefore, this makes implementation easy.

Akallouch et al. (2020) attempted to build a dataset which can be used by models that extracts text from traffic panels with Arabic scripts, for which very few datasets exist [59]. The authors called this dataset ASAYAR, and it is made up of three sub-datasets: directional symbol detection, traffic sign detection, and Arabic-Latin text localization. The dataset consists of around 1763 images. This was utilized by various models after it. Kim et al. (2020) [60] suggested a Generative Adversarial Network based Text Localization model that they called TLGAN. It detects the text location by translating an image into a text localization map, and then finding text-bounding boxes from the map. This model achieved a precision score of 99.83%, and a recall score of 99.64%. The following table lists some of the papers published in the last five years, discussing text localization and recognition in images.

Table 1. List of published papers in the last five years with novel approaches for text localization and recognition in images.

Author, Year	Methods used for pre-processing/image-processing	Dataset	Classifiers / ML Models used	Inference
Chandio *et al.* (2018) [61]	Histogram of Oriented Gradient (HOG) method	Manually segmented characters – 18000 cropped Urdu characters	SVM, kNN, Random Forest, ExtraTree Classifier, and MLP	The authors presented a functional method to recognize and classify Urdu text in images. The highest accuracy is attained by a 10-fold ExtraTree Model.
Dutta *et al.* (2019) [62]	Binning, Connected Component Analysis (CCA)	In-house dataset of 300+ smartphone camera-	-	The proposed method converts colour images

		captured images.		into binary images, and then subjects it to various elimination methods until the potential text regions are detected and located. The model is language independent and works for several languages such as Oriya, Bangla, Hindi, and English, achieving a best case F-measure value of 0.69.
Khan *et al.* (2018) [63]	Background Suppression, Connected Component Analysis (CCA)	227 camera-captured images	MLP	The proposed model is trained and testes on a dataset of 227 camera-captured images, and the accuracy score for object isolation hence obtained is 0.8638, and the accuracy score for classification

				of text/non-text is 0.9648.
Paula*etaal.* (2019) [64]	Adaptive Stroke Filter based on Fuzzy Distance Transform	600 synthetic, mobile-captured, and ICDAR 2011 images	-	The proposed model attains a recall score of 96.65%, a precision score of 87.77%, and an f-measure score of 91.89%. The stroke filter does not work well for non-binarized images with varying font sizes and has very high time complexity. The proposed model solves these issues because it works for grayscale images, dynamically designs the filter on the basis of the FDT value, and has very low time complexity.
Roopa *et al.* (2019) [65]	Sharpening, Maximally Stable Extremal Regions (MSER) detector	ICDAR 2015	-	After enhancing the images of the ICDAR 2015 incidental dataset, the

				model discussed in the paper achieves a precision and recall score of 54.61% and 64.29% respectively.

IV. VIDEO TEXT RECOGNITION

Text discovery in videos and images has several benefits. For one, it may be used to make Social Media platforms that are mostly based on sharing photos and videos more accessible for the visually-impaired[66]. The text in an image, meme, or a YouTube video can be a useful parameter in NLP-based recommendation algorithms if it does not have effective tags. It also has various other real-world applications such as traffic-sign/signboard detection, subtitling and captioning, and transcripts and notes for educational videos.

Among the several approaches for text localization and detection in images and videos are those based on wavelet features, clustering and segmentation approaches, edge detection techniques, motion-based detection, colour-based approaches, gradient-based techniques, approaches based on neural networks, SVM-based methods, Independent Component Analysis algorithms, morphology-based approaches, methods based on corner detection, texture analysis, compression-based methods, Laplacian approaches, models based on tracking and prediction etc. This isn't an exhaustive list as an extensive body of research on text recognition exists and is growing exponentially [57].

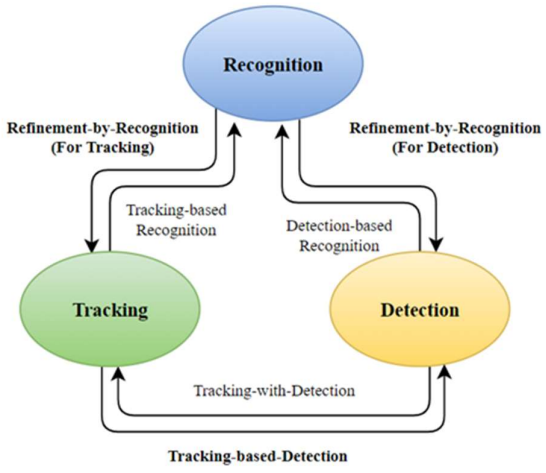

Fig.5. Direction-Tracking-Recognition framework for video text detection.

The unified video text DEtection, Tracking and Recognition (DETR) framework is shown in Fig.2. Here,This framework uniformly describes detection, tracking, recognition (the three main tasks), and their relations and interactions. The major relations among these main tasks are unified as "detection-based recognition", "tracking-based-detection" and "tracking-based-

recognition". The other relations among these tasks are also named as "refinement-by recognition (for detection)", "refinement-by-recognition (for tracking)" and "tracking-with-detection".

Many recent works have projected each iteration process into an output sequence of varying lengths. This is done with the assumption that scene text is typically consisting of a list of a sequence of characters. Several sequence-based text recognition systems [67-73] have been using connectionist temporal classification [74] enabling estimation of character sequences. These methods were inspired by the challenge of speech recognition. For the first case [75], CNN models have been combined with CTC. Examples of these models include VGG [76], RCNN [70], and ResNet[77]. For example, in Yin iet al. (2017), a sliding window is initially employed to the picture to efficiently collect appropriate data, and then a CTC extrapolation is employed in order to forecast the outcome of the words [75].

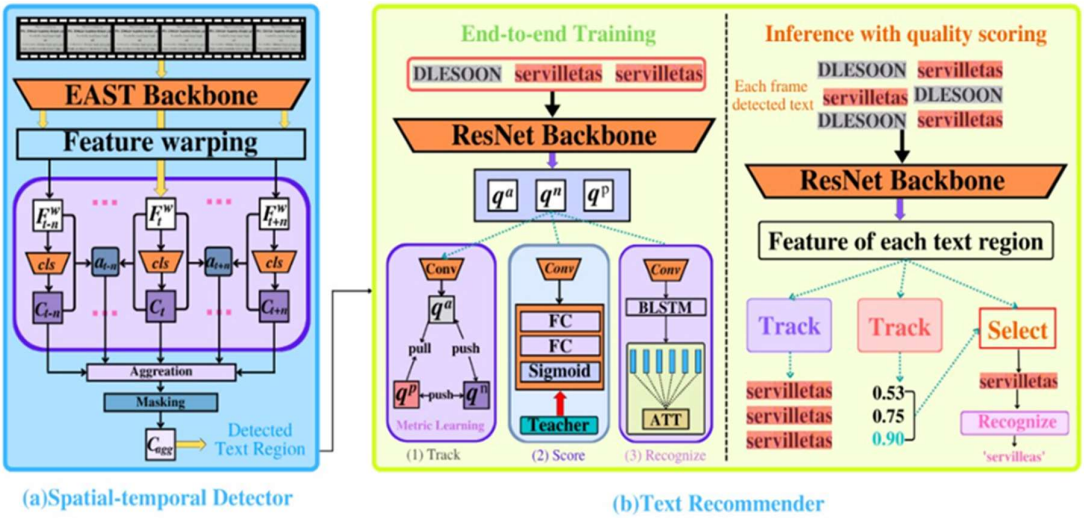

Fig.6. The workflow of YORO (You Only Recognize Once) system.

YORO consists of two modules namely spatial temporal text detecting to identify text regions and recommender to assemble the quality and track. The left part is called as training process and right part is called as inference model. Here attention mechanism will be applied respectively.

Cheng et al. (2021) [78] proposed a video text-spotting model that they called FREE. FREE stands for Fast and Robust End-to-End Video Text Spotter. FREE uses a spatiotemporal detector that learns text locations among video frames. The dataset consists of 100 videos collected from across 21 real-world scenarios and is called the Large-Scale Video Text Dataset or LSVTD. FREE attains better f1 scores for Detection and End-to-End tasks as compared to other models like EAST+CRNN on the LSVTD
dataset.
Mirza et al. (2020) present a framework for the detection of cursive Urdu text in videos [79]. This model is based on CNN for script identification of detected textual content, and on

CNN+LSTM for recognition. A dataset of 13000+ video frames is compiled and the model achieves an f-measure score of 88.3% and an 87% recognition rate.
Rosetta system (2018) made use of merely the extracted features from the convolutional neural network by employing a ResNet model as a backbone in order to make predictions about the feature sequences [69]. These approaches [69, 75] suffered from a deficiency of relevant data and demonstrated a low recognition accuracy, despite the fact that they reduced the complexity of the computational work. Cheng et al. (2021) developed a technique for rapid and reliable video text spotting. Their method consisted of just detecting the localised text a single time, as opposed to identifying it frame-by-frame.

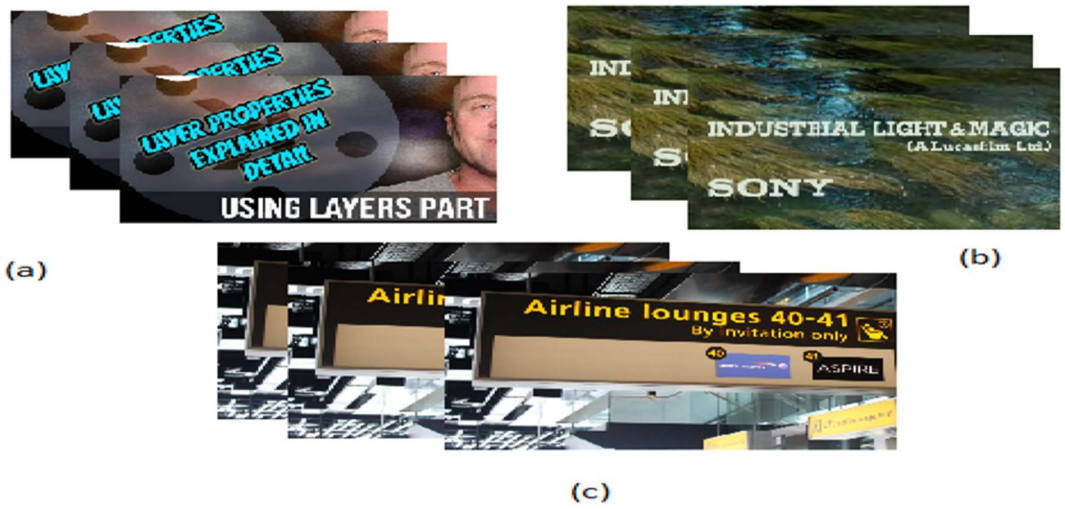

Fig. 7.Types of texts present in videos - (a) Layered caption text, (b) embedded caption text, and (c) scene text, where embedded caption text is common and challenging for detection, tracking and recognition from web videos.

The deep learning approach in Mirza et al. (2020) uses CNN and RNN method to work on cursive lettered-image input. This work created a labelled video frame data set. Video frames were analysed to derive text line images. Preprocessed extracted lines were used to train CNN–RNN end-to-end. A series of experiments were conducted to examine recognition performance based on pre-processing, RNN cell type, and training data. The system recognised 97.63% of text-line pictures [80].

The following table lists some of the papers published in the last five years, discussing text localization and recognition in videos.

Table 2. List of published papers in the last five years with novel approaches for text localization and recognition in videos.

Author, Year	Methods Used for pre-processing/image processing	Dataset	Classifiers / ML Models used	Inference

Cai *et al.* (2018) [81]	Sampling & Recovery Model (SaRM), Divide & Conquer Model (DaCM)	UCAS-STL Dataset with 57070 frames	-	The spatiotemporal text detection model includes SaRM and DaCM achieves impressive results when compared to edge-based, SWT-based, and Deep-CNN based approaches.
Dutta *et al.* (2018) [82]	EAST and Textboxes++	LectureVideoDB, IAM Handwriting Database, MJSynth	Hybrid CNN+RNN – CRNN, CRNN-STN	The proposed model seeks to extract text from Lecture Videos that use slides, whiteboards, paper, or blackboards. The model shows impressive performance based on query-by-string (QBS) as well as query-by-example (QBE). It achieves a mean average precision value of 0.7909 and 0.7404 in with

				respect to QBE and QBS respectively.
Reddy *et al.* (2020) [83]	-	Novel RoadText-1K dataset	CRNN-ASTER	The novel RoadText-1K dataset is 20 times larger than other datasets for driving videos. It consists of 4% License Plate text instances, 65.7% English, 28.3% non-English, and 2% illegible text instances. RoadText-1K concerns itself with real-world text-spotting instances like capturing text from a moving car.
Yang *et al.* (2018) [84]	ResNet-50 and PVANET	Dataset I (comprises of 15 videos with English captions) Dataset II (comprises of 15 videos with Chinese captions) Dataset III (comprises of 15 bilingual video sequences)	Fully Convolutional Networks	The proposed model shows high performance in real-world video datasets and can detect English or Chinese text without word segmentation. The output of this model can then be used

				for text recognition.
Yin *et al.* (2018) [42]	Grayscale, Sobel Operator-based Edge detection	300 video images from across the internet, containing mostly Chinese and some English characters	Adaboost, CART (Classification and Regression Tree)	The Adaboost based strong classifier formed by five weak features has a Precision Score of 0.761 and a Recall Score of 0.740. This model shows evidently better performance than just a plain SVM model.

V. CONCLUSION

Text that is included in a video and images offers a distinct and helpful source of information on the video and images' subject matter. Retrieval of this information for semantic indexing is made possible by the localisation of text in films, segmentation of that text, and recognition of that text. However, due to the disparity between the current technological position and the essential enactment, it is clear that text detection and identification are still disputes that need to be resolved. Despite the significant advancements that have been accomplished, there are still a great many prospects for research. A few prospects for research can be summarized as follows:

End-to-end text:

End-to-end text recognition accuracy is still far behind when compared to OCR's performance on clean documents [85-88]. Sturdier character recognition architectures will help, but so will well-planned information sharing, feedback, and optimization procedures.

Analyzing incidental text:

Incidental text is harmed by background clutter [89], picture deterioration [90], distortions [91], and font variations [92-94]. Many strategies can address a particular issue, but very few strategies really combine them. Improved invariant characteristics must be constructed or learnt, cutting-edge augmentation and rectification techniques must be used, and new sensors must be used to solve the broader issue of accidental text detection and identification.

Detection of text from different languages:

Text from different languages has distinct features while being processed in a multilingual manner [94-96]. Due to the numerous character classes, intricate character structures, similarity of characters, and variety of typefaces, the recognition of text from East Asian nations such as China, Japan, and Korea (CJK) was thought to be an incredibly challenging task[97-99]. It is still challenging to recognize text from all languages using a single approach with set parameters. One approach is to provide a model for each type of language using a common trainable technique and to manage the models using a customizable manner.

REFERENCES

[1] H. D. Wactlar, T. Kanade, M. A. Smith, and S. M. J. C. Stevens, "Intelligent access to digital video: Informedia project," vol. 29, no. 5, pp. 46-52, 1996.

[2] B. W. Wojdynski and N. J. J. J. o. A. Evans, "Going native: Effects of disclosure position and language on the recognition and evaluation of online native advertising," vol. 45, no. 2, pp. 157-168, 2016.

[3] C. Reis, A. Moshchuk, and N. Oskov, "Site isolation: Process separation for web sites within the browser," in 28th USENIX Security Symposium (USENIX Security 19), 2019, pp. 1661-1678.

[4] S.-X. Zhang et al., "Deep relational reasoning graph network for arbitrary shape text detection," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 9699-9708.

[5] K. A. Al-Thelaya, T. S. Al-Nethary, and E. Y. Ramadan, "Social Networks Spam Detection Using Graph-Based Features Analysis and Sequence of Interactions Between Users," in 2020 IEEE International Conference on Informatics, IoT, and Enabling Technologies (ICIoT), 2020, pp. 206-211: IEEE.

[6] S. Khalandi, F. J. J. o. A. i. C. E. Soleimanian Gharehchopogh, and Technology, "A new approach for text documents classification with invasive weed optimization and naive bayes classifier," vol. 4, no. 3, pp. 167-184, 2018.

[7] A. I. J. A. I. R. Kadhim, "Survey on supervised machine learning techniques for automatic text classification," vol. 52, no. 1, pp. 273-292, 2019.

[8] Z.-R. Wang, J. Du, W.-C. Wang, J.-F. Zhai, J.-S. J. I. J. o. D. A. Hu, and Recognition, "A comprehensive study of hybrid neural network hidden Markov model for offline handwritten Chinese text recognition," vol. 21, no. 4, pp. 241-251, 2018.

[9] B. Mor, S. Garhwal, and A. J. A. o. c. m. i. e. Kumar, "A systematic review of hidden Markov models and their applications," vol. 28, no. 3, pp. 1429-1448, 2021.

[10] L. Yang, P. Wang, H. Li, Z. Li, and Y. J. N. Zhang, "A holistic representation guided attention network for scene text recognition," vol. 414, pp. 67-75, 2020.

[11] D. Li, C. Rodriguez, X. Yu, and H. Li, "Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison," in Proceedings of the IEEE/CVF winter conference on applications of computer vision, 2020, pp. 1459-1469.

[12] D. Das, D. R. Nayak, R. Dash, B. Majhi, and Y. D. J. I. I. P. Zhang, "H-WordNet: a holistic convolutional neural network approach for handwritten word recognition," vol. 14, no. 9, pp. 1794-1805, 2020.

[13] P. Lyu, C. Yao, W. Wu, S. Yan, and X. Bai, "Multi-oriented scene text detection via corner localization and region segmentation," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7553-7563.
[14] O. Zayene, S. M. Touj, J. Hennebert, R. Ingold, and N. E. J. I. C. V. Ben Amara, "Multi-dimensional long short-term memory networks for artificial Arabic text recognition in news video," vol. 12, no. 5, pp. 710-719, 2018.
[15] O. Zayene, S. Masmoudi Touj, J. Hennebert, R. Ingold, and N. J. J. o. I. Essoukri Ben Amara, "Open datasets and tools for arabic text detection and recognition in news video frames," vol. 4, no. 2, p. 32, 2018.
[16] A. Mahajan, A. Nayyar, R. Jain, and P. Nagrath, "Natural Scenes' Text Detection and Recognition Using CNN and Pytesseract," in The Fifth International Conference on Safety and Security with IoT, 2023, pp. 159-171: Springer.
[17] Y. Xu et al., "End-to-end subtitle detection and recognition for videos in East Asian languages via CNN ensemble," vol. 60, pp. 131-143, 2018.
[18] K. I. Kim, K. Jung, J. H. J. I. T. o. P. A. Kim, and M. Intelligence, "Texture-based approach for text detection in images using support vector machines and continuously adaptive mean shift algorithm," vol. 25, no. 12, pp. 1631-1639, 2003.
[19] K. Wang, B. Babenko, and S. Belongie, "End-to-end scene text recognition," in 2011 International conference on computer vision, 2011, pp. 1457-1464: IEEE.
[20] J.-J. Lee, P.-H. Lee, S.-W. Lee, A. Yuille, and C. Koch, "Adaboost for text detection in natural scene," in 2011 International conference on document analysis and recognition, 2011, pp. 429-434: IEEE.
[21] L. Neumann and J. Matas, "A method for text localization and recognition in real-world images," in Asian conference on computer vision, 2010, pp. 770-783: Springer.
[22] C. Yao, X. Bai, W. Liu, Y. Ma, and Z. Tu, "Detecting texts of arbitrary orientations in natural images," in 2012 IEEE conference on computer vision and pattern recognition, 2012, pp. 1083-1090: IEEE.
[23] L. Neumann and J. Matas, "Scene text localization and recognition with oriented stroke detection," in Proceedings of the ieee international conference on computer vision, 2013, pp. 97-104.
[24] X. Liu, G. Meng, C. J. I. J. o. D. A. Pan, and Recognition, "Scene text detection and recognition with advances in deep learning: a survey," vol. 22, no. 2, pp. 143-162, 2019.
[25] A. Krizhevsky, I. Sutskever, and G. E. J. A. i. n. i. p. s. Hinton, "Imagenet classification with deep convolutional neural networks," vol. 25, 2012.
[26] B. Shi, X. Bai, and S. Belongie, "Detecting oriented text in natural images by linking segments," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2550-2558.
[27] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. J. I. j. o. c. v. Zisserman, "Reading text in the wild with convolutional neural networks," vol. 116, no. 1, pp. 1-20, 2016.
[28] W. Huang, Y. Qiao, and X. Tang, "Robust scene text detection with convolution neural network induced mser trees," in European conference on computer vision, 2014, pp. 497-511: Springer.

[29] Z. Zhang, W. Shen, C. Yao, and X. Bai, "Symmetry-based text line detection in natural scenes," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 2558-2567.
[30] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. J. a. p. a. Zisserman, "Deep structured output learning for unconstrained text recognition," 2014.
[31] J. Matas, O. Chum, M. Urban, T. J. I. Pajdla, and v. computing, "Robust wide-baseline stereo from maximally stable extremal regions," vol. 22, no. 10, pp. 761-767, 2004.
[32] P. Dollár, R. Appel, S. Belongie, P. J. I. t. o. p. a. Perona, and m. intelligence, "Fast feature pyramids for object detection," vol. 36, no. 8, pp. 1532-1545, 2014.
[33] J. Ma et al., "Arbitrary-oriented scene text detection via rotation proposals," vol. 20, no. 11, pp. 3111-3122, 2018.
[34] M. Liao, B. Shi, X. Bai, X. Wang, and W. Liu, "Textboxes: A fast text detector with a single deep neural network," in Thirty-first AAAI conference on artificial intelligence, 2017.
[35] W. Liu et al., "Ssd: Single shot multibox detector," in European conference on computer vision, 2016, pp. 21-37: Springer.
[36] Z. Tian, W. Huang, T. He, P. He, and Y. Qiao, "Detecting text in natural image with connectionist text proposal network," in European conference on computer vision, 2016, pp. 56-72: Springer.
[37] S. Ren, K. He, R. Girshick, and J. J. A. i. n. i. p. s. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," vol. 28, 2015.
[38] A. Gupta, A. Vedaldi, and A. Zisserman, "Synthetic data for text localisation in natural images," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2315-2324.
[39] J. Redmon, S. Divvala, R. Girshick, and A. J. P. Farhadi, NJ: IEEE., "You only look once: Unified, real-time object detection In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 779–788)," 2016.
[40] R. Lienhart, A. J. I. T. o. c. Wernicke, and s. f. v. technology, "Localizing and segmenting text in images and videos," vol. 12, no. 4, pp. 256-268, 2002.
[41] P. He, W. Huang, T. He, Q. Zhu, Y. Qiao, and X. Li, "Single shot text detector with regional attention," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 3047-3055.
[42] F. Yin, R. Wu, X. Yu, G. J. M. T. Sun, and Applications, "Video text localization based on Adaboost," vol. 78, no. 5, pp. 5345-5354, 2019.
[43] F. Zhan and S. Lu, "ESIR: end-to-end scene text recognition via iterative rectification," in Proceedings of IEEE conference on Computer Vision and Pattern Recognition, 2019. 2059, vol. 2068.
[44] J. Lafferty, A. McCallum, and F. C. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," 2001.
[45] M. Li and C. Wang, "An adaptive text detection approach in images and video frames," in 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), 2008, pp. 72-77: IEEE.
[46] M. Jaderberg, A. Vedaldi, and A. Zisserman, "Deep features for text spotting," in European conference on computer vision, 2014, pp. 512-528: Springer.

[47] Y. Wang, H. Xie, Z. Zha, Y. Tian, Z. Fu, and Y. J. I. T. o. M. Zhang, "R-Net: A relationship network for efficient and accurate scene text detection," vol. 23, pp. 1316-1329, 2020.
[48] Y. Wu and P. Natarajan, "Self-organized text detection with minimal post-processing via border learning," in proceedings of the IEEE international conference on computer vision, 2017, pp. 5000-5009.
[49] Q. Ye, W. Gao, W. Wang, and W. Zeng, "A robust text detection algorithm in images and video frames," in Fourth International Conference on Information, Communications and Signal Processing, 2003 and the Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint, 2003, vol. 2, pp. 802-806: IEEE.
[50] A. Mishra, K. Alahari, and C. Jawahar, "Scene text recognition using higher order language priors," in BMVC-British machine vision conference, 2012: BMVA.
[51] L. Neumann and J. Matas, "Real-time scene text localization and recognition," in 2012 IEEE conference on computer vision and pattern recognition, 2012, pp. 3538-3545: IEEE.
[52] T. Wang, D. J. Wu, A. Coates, and A. Y. Ng, "End-to-end text recognition with convolutional neural networks," in Proceedings of the 21st international conference on pattern recognition (ICPR2012), 2012, pp. 3304-3308: IEEE.
[53] C. Yao, X. Bai, and W. J. I. T. o. I. P. Liu, "A unified framework for multioriented text detection and recognition," vol. 23, no. 11, pp. 4737-4749, 2014.
[54] M. Busta, L. Neumann, and J. Matas, "Deep textspotter: An end-to-end trainable scene text localization and recognition framework," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 2204-2212.
[55] T. He, Z. Tian, W. Huang, C. Shen, Y. Qiao, and C. Sun, "An end-to-end textspotter with explicit alignment and attention," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 5020-5029.
[56] X. Chen, L. Jin, Y. Zhu, C. Luo, and T. J. A. C. S. Wang, "Text recognition in the wild: A survey," vol. 54, no. 2, pp. 1-35, 2021.
[57] V. Manjunath Aradhya, H. Basavaraju, and D. S. J. E. I. Guru, "Decade research on text detection in images/videos: a review," vol. 14, no. 2, pp. 405-431, 2021.
[58] K. Raghunandan et al., "Multi-script-oriented text detection and recognition in video/scene/born digital images," vol. 29, no. 4, p. 1145, 2019.
[59] M. Akallouch, K. S. Boujemaa, A. Bouhoute, K. Fardousse, and I. J. I. T. o. I. T. S. Berrada, "ASAYAR: A dataset for Arabic-Latin scene text localization in highway traffic panels," 2020.
[60] D. Kim, M. Kwak, E. Won, S. Shin, and J. J. a. p. a. Nam, "TLGAN: document Text Localization using Generative Adversarial Nets," 2020.
[61] A. A. Chandio, M. Pickering, and K. Shafi, "Character classification and recognition for Urdu texts in natural scene images," in 2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET), 2018, pp. 1-6: IEEE.
[62] I. N. Dutta, N. Chakraborty, A. F. Mollah, S. Basu, and R. Sarkar, "Multi-lingual text localization from camera captured images based on foreground homogenity analysis," in Recent developments in machine learning and data analytics: Springer, 2019, pp. 149-158.

[63] T. Khan and A. F. Mollah, "A novel text localization scheme for camera captured document images," in Proceedings of 2nd International Conference on Computer Vision & Image Processing, 2018, pp. 253-264: Springer.
[64] S. Paul, S. Saha, S. Basu, P. K. Saha, M. J. M. T. Nasipuri, and Applications, "Text localization in camera captured images using fuzzy distance transform based adaptive stroke filter," vol. 78, no. 13, pp. 18017-18036, 2019.
[65] M. Roopa and K. Mahantesh, "An impact of frequency domain filtering technique on text localization method useful for text reading from scene images," in 2019 4th International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques (ICEECCOT), 2019, pp. 37-43: IEEE.
[66] V. Sivakumar, A. Gordo, and M. J. F. E. b. p. o. Paluri, "Rosetta: Understanding text in images and videos with machine learning," vol. 11, p. 2018, 2018.
[67] B. Shi, X. Bai, C. J. I. t. o. p. a. Yao, and m. intelligence, "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition," vol. 39, no. 11, pp. 2298-2304, 2016.
[68] W. Liu, C. Chen, K.-Y. K. Wong, Z. Su, and J. Han, "Star-net: a spatial attention residue network for scene text recognition," in BMVC, 2016, vol. 2, p. 7.
[69] F. Borisyuk, A. Gordo, and V. Sivakumar, "Rosetta: Large scale system for text detection and recognition in images," in Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining, 2018, pp. 71-79.
[70] C.-Y. Lee and S. Osindero, "Recursive recurrent nets with attention modeling for ocr in the wild," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2231-2239.
[71] J. Wang and X. J. A. i. N. I. P. S. Hu, "Gated recurrent convolution neural network for ocr," vol. 30, 2017.
[72] Y. Liu, Z. Wang, H. Jin, and I. Wassell, "Synthetically supervised feature learning for scene text recognition," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 435-451.
[73] P. He, W. Huang, Y. Qiao, C. C. Loy, and X. Tang, "Reading scene text in deep convolutional sequences," in Thirtieth AAAI conference on artificial intelligence, 2016.
[74] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in Proceedings of the 23rd international conference on Machine learning, 2006, pp. 369-376.
[75] F. Yin, Y.-C. Wu, X.-Y. Zhang, and C.-L. J. a. p. a. Liu, "Scene text recognition with sliding convolutional character models," 2017.
[76] K. Simonyan and A. J. a. p. a. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014.
[77] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770-778.
[78] Z. Cheng et al., "Free: A fast and robust end-to-end video text spotter," vol. 30, pp. 822-837, 2020.

[79] A. Mirza, O. Zeshan, M. Atif, I. J. E. J. o. I. Siddiqi, and V. Processing, "Detection and recognition of cursive text from video frames," vol. 2020, no. 1, pp. 1-19, 2020.
[80] A. Mirza and I. J. I. I. P. Siddiqi, "Recognition of cursive video text using a deep learning framework," vol. 14, no. 14, pp. 3444-3455, 2020.
[81] Y. Cai, W. Wang, S. Huang, J. Ma, K. J. M. T. Lu, and Applications, "Spatiotemporal text localization for videos," vol. 77, no. 22, pp. 29323-29345, 2018.
[82] K. Dutta, M. Mathew, P. Krishnan, and C. Jawahar, "Localizing and recognizing text in lecture videos," in 2018 16th international conference on frontiers in handwriting recognition (ICFHR), 2018, pp. 235-240: IEEE.
[83] S. Reddy, M. Mathew, L. Gomez, M. Rusinol, D. Karatzas, and C. Jawahar, "Roadtext-1k: Text detection & recognition dataset for driving videos," in 2020 IEEE International Conference on Robotics and Automation (ICRA), 2020, pp. 11074-11080: IEEE.
[84] C. Yang, W.-Y. Pei, L.-H. Wu, and X.-C. J. B. D. A. Yin, "Chinese text-line detection from web videos with fully convolutional networks," vol. 3, no. 1, pp. 1-11, 2018.
[85] F. Makhmudov et al., "Improvement of the end-to-end scene text recognition method for “text-to-speech” conversion," vol. 18, no. 06, p. 2050052, 2020.
[86] S. Qin, A. Bissacco, M. Raptis, Y. Fujii, and Y. Xiao, "Towards unconstrained end-to-end text spotting," in Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 4704-4714.
[87] S. Das et al., "End-to-end Piece-wise Unwarping of Document Images," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 4268-4277.
[88] Y. Deng, D. Rosenberg, and G. Mann, "Challenges in end-to-end neural scientific table recognition," in 2019 International Conference on Document Analysis and Recognition (ICDAR), 2019, pp. 894-901: IEEE.
[89] T. Guan et al., "Industrial Scene Text Detection with Refined Feature-attentive Network," 2022.
[90] T. Küstner, K. Armanious, J. Yang, B. Yang, F. Schick, and S. J. M. r. i. m. Gatidis, "Retrospective correction of motion-affected MR images using deep learning frameworks," vol. 82, no. 4, pp. 1527-1540, 2019.
[91] X. Liu, D. Liang, S. Yan, D. Chen, Y. Qiao, and J. Yan, "Fots: Fast oriented text spotting with a unified network," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 5676-5685.
[92] S. Cahyadi, F. Damatraseta, and V. I. Sugara, "Comparative Analysis Of Efficient Image Segmentation Technique For Text Recognition And Human Skin Recognition," in IOP Conference Series: Materials Science and Engineering, 2019, vol. 621, no. 1, p. 012007: IOP Publishing.
[93] Y. Liu, L. Jin, and C. J. I. T. o. I. P. Fang, "Arbitrarily shaped scene text detection with a mask tightness text detector," vol. 29, pp. 2918-2930, 2019.
[94] A. Giri, "An Efficient Preprocessing Module For Incidental Scene Text Recognition," Indian Statistical Institute-Kolkata, 2020.
[95] C. Jawahar, "Transfer Learning for Scene Text Recognition in Indian Languages."

[96] W. He, X.-Y. Zhang, F. Yin, Z. Luo, J.-M. Ogier, and C.-L. J. P. R. Liu, "Realtime multi-scale scene text detection with scale-based region proposal network," vol. 98, p. 107026, 2020.
[97] N. Nayef et al., "ICDAR2019 robust reading challenge on multi-lingual scene text detection and recognition—RRC-MLT-2019," in 2019 International conference on document analysis and recognition (ICDAR), 2019, pp. 1582-1587: IEEE.
[98] W. He, X.-Y. Zhang, F. Yin, and C.-L. J. I. T. o. I. P. Liu, "Multi-oriented and multi-lingual scene text detection with direct regression," vol. 27, no. 11, pp. 5406-5419, 2018.
[99] S.-G. Lee, Y. Sung, Y.-G. Kim, and E.-Y. J. J. o. i. p. s. Cha, "Variations of AlexNet and GoogLeNet to improve Korean character recognition performance," vol. 14, no. 1, pp. 205-217, 2018.