The Academic
International Journal of Multidisciplinary Research

An Online Peer Reviewed / Refereed Journal
Volume 1 | Issue 2 | July-September 2023
ISSN: 2583-973X
Website: www.theacademic.in

Comparison of Various Approaches for Real Time Multilingual Text and Object Detection in Images

Pranav Rajendra Patil*
Research Scholar
KCES's, M. J. College, Jalgaon, Maharashtra, India

Dr. Monali Y. Khachane
Asst. Professor
Dept. of Computer Science, Dr. Annasaheb G.D. Bendale Mahila Mahavidhyalaya, Jalgaon, Maharashtra, India

ARTICLE DETAILS

Research Paper

Article History
Received : September 09, 2023
Accepted : September 23, 2023

Keywords :
Text Identification, Image Processing, OCR, Mask RCNN, RCNN, Artificial Neural Networks, Machine Learning

ABSTRACT

Ever since the advent of artificial intelligence, there has been a keen interest among developers in imbuing computers with human-like thinking capabilities. The conception of Artificial Neural Networks stemmed from the ambition to facilitate this transition. These sophisticated machines find applications in diverse fields like robotics, medicine, and industry. With the emergence of image recognition algorithms pioneered by Facebook's AI researchers, image identification and recognition have captured the imagination of developers. The primary aim of image identification software is to discern the elements within an image or scene and distinguish them. Various algorithms such as OCR, RCNN, Mask RCNN, Fast RCNN, and Faster RCNN have been devised to achieve this, classifying images into categories as dictated by the programmer's intent. This paper delves into an exploration of the aforementioned Image Processing algorithms, seeking to ascertain their respective strengths. Research has shown that Mask RCNN, an enhanced iteration of Faster RCNN, has demonstrated itself as the most adept algorithm in the realm of real-time object detection.

*Corresponding Author

1. INTRODUCTION

Real-world applications are increasingly seeking to leverage computers to streamline tasks for people across various domains. Owing to their user-friendly interfaces, computers have found widespread utility in fields spanning from robotics and medicine to advanced computing. Similarly, when robots are furnished with data about their surroundings, they are better equipped to comprehend their immediate environments and respond effectively to different situations. Likewise, if a computer is trained to discern and identify tumor cells, it can significantly expedite the process of tumor and cancer cell identification, alleviating the manual labor involved in this task [1]. In the case of self-driving cars, it is imperative for the system to not only identify objects in still images but also in real time, enabling prompt and appropriate responses.

Object detection seeks to imbue systems with the ability to understand visual patterns within an image [2]. The primary goal of this line of study is to master the modeling of various deformities, inclusions, and other class-specific variations while efficiently handling extensive datasets across diverse conditions.

Machine Learning (ML) [3], a subset of artificial intelligence, is dedicated to training machines and crafting algorithms to manage tasks such as robotics. These machines learn to operate autonomously after being trained on extensive datasets. Rather than being explicitly programmed for a single specialized task, they are equipped to perform a multitude of tasks concurrently. The self-iterative nature of these algorithms allows them to learn and enhance themselves. When confronted with new input data, the ML algorithm makes predictions based on the model it has developed. These predictions are then assessed for accuracy, and if deemed acceptable, the Machine Learning algorithm is deployed. If the accuracy falls below the acceptable threshold, the algorithm undergoes further iterations, often with an expanded training dataset, to improve its performance.

The pivotal question that arises is: how can machines be trained to emulate human thought processes? This question finds its answer in the realm of artificial neural networks (ANNs). ANNs [5] serve as invaluable tools, particularly in the domain of artificial intelligence, especially in tasks like facial recognition, where they help mitigate errors stemming from human intervention. As the name implies, ANNs function as proficient computing systems with a structure akin to that of parallel distributed processing and connectionist systems. They process a large dataset collectively, which is then

transmitted through a network of interconnected units, commonly referred to as nodes or neurons, to yield the desired output.

In contrast, neural networks [6] are adept at assisting individuals in solving complex real-world problems. They possess the capability to learn and establish relationships among inputs and outputs that are non-linear and intricate. Additionally, they excel at generalizing, inferring, revealing hidden correlations, patterns, and predictions, and generating highly dynamic data, including financial time series, along with the variances necessary for predicting exceedingly rare events, such as fraud detection. Neurons are linked to one another through specialized connections known as links.

These neurons are predominantly utilized in the realm of facial recognition software, giving rise to a new branch of image processing. Image processing entails the manipulation of images through a computer system. Much like how a computer processes numerical data, in this case, computers are tasked with handling images instead of raw data. Its objective is to create a computer algorithm that performs operations on images. The input is a digital image, and the program processes it using predefined algorithms, ultimately generating an output in the form of an enhanced image.

2. LITERATURE REVIEW

A more contemporaneous stride in the advancement of artificial neural networks was pioneered by the PDP group. They introduced a neural-based classification system, employing Parallel Hopfield Neural Networks for neutral facial image recognition. In 2015, Alexandrina-Elena Pandelea [5] et al. harnessed image processing within neural networks for geotechnical engineering, successfully mitigating landslides by training the network with ASTER images and GIS, followed by a subsequent generation of learning maps.

Minghui Liao [7] et al. presented the "text box" approach in 2016 for swift scene detection. This end-to-end model demonstrated rapid scene detection capabilities, achieving high accuracy and precision within a single network, without necessitating post-processing of the image—apart from a standard non-maximum suppression step. This method exhibited exceptional speed in text detection, clocking in at approximately 0.09 seconds, a remarkably swift performance for its time.

In 2017, the Mask-RCNN model emerged as an extension of the Faster-RCNN model, specifically designed for semantic division, object localization, and instance division within natural images. Mask-RCNN outperformed all preceding models comprehensively, as evidenced by its dominant performance

in the 2016 COCO Challenge [8], a large-scale endeavor encompassing object detection, segmentation, and captioning.

Kaiming He [9] et al. introduced, in 2018, a notably straightforward, adaptable, and streamlined approach to object case segmentation. Their technique exhibited remarkable accuracy in object detection, coupled with the generation of high-fidelity segmentation masks for each occurrence. Results were meticulously documented across various metrics including instance segmentation, bounding box object detection, and key point detection of individuals in the COCO challenge. Without a doubt, Mask R-CNN outshone all existing single-model entries across every task, even surpassing the winners of the COCO 2016 challenge.

The utilization of HOG and SIFT [10] pyramids has been prevalent in numerous studies encompassing instance classification, object detection, human pose estimation, and more. Dollár [11] et al. introduced a rapid pyramid computation method by initially generating a sparsely sampled scale pyramid, and then integrating the missing layers. Prior to HOG and SIFT, initial endeavors in appearance recognition with ConvNets involved the computation of narrow systems over image pyramids to detect features across varying scales. Recent advances encompass models like ResNet, Inception-ResNet, and ResNetXt, tailored specifically for object detection.

2.1. Mask RCNN

Mask RCNN stands as the foremost cutting-edge deep learning algorithm, excelling in object detection and instance segmentation. Its rapid ascent in popularity is the central focus of our review. Mask RCNN [13] has demonstrated remarkable performance on the MSCOCO dataset [14]. The process behind Mask-RCNN involves a two-stage image recognition approach. Firstly, it extracts features from the image utilizing a backbone Convolutional Neural Network (CNN) and predicts class-agnostic regional proposals. These proposals are subsequently refined and organized in the second stage to form either labeled bounding boxes for object detection or segmentation masks for instance segmentation tasks.

This algorithm is adept at discerning distinct objects within an image or video. Much of its progress has been propelled by influential precursor systems, namely the Fast/Faster RCNN [11], and the Fully Convolutional Network (FCN) [15] architectures for object detection and semantic segmentation, respectively. The objective is to introduce an equally enabling framework for instance segmentation. Instance segmentation [16] is particularly challenging as it necessitates the precise identification of all

objects in an image while also accurately delineating each occurrence. Additionally, single-shot models like YOLO and SSD have revolutionized object detection, achieving speeds up to 100-1000 times faster than area proposal-based algorithms.

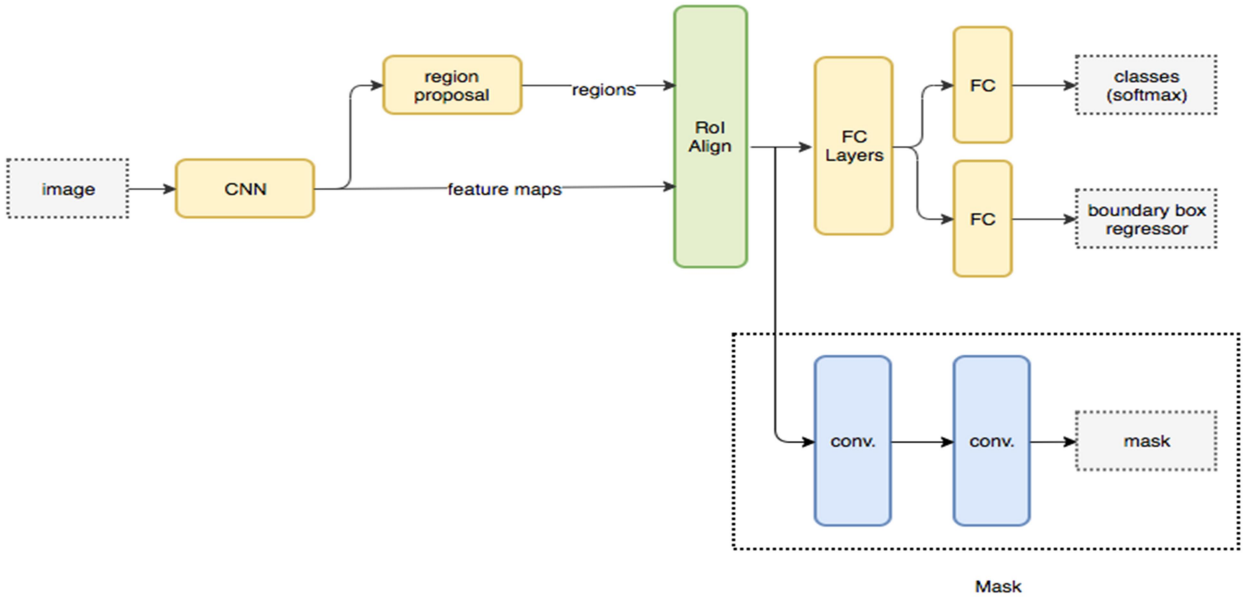

Figure 1: The Mask R-CNN framework for instance segmentation [12]

Mask R-CNN [12], named for the object mask appended to object class labels with the inclusion of bounding boxes, represents an enhanced version of Fast RCNN. It entails the meticulous mining of finer pixels within an object. Moreover, Mask RCNN places greater emphasis on pixel-to-pixel alignment in an object's layout, a crucial component that was lacking in both Fast and Faster RCNN.

2.2. RoIAlign

RoIPool is a relatively straightforward feature map. Its operation involves the initial quantization of the entire spatial layout into distinct granules, followed by a further division of these granules into spatial bins. This process results in a finely partitioned spatial plane. Subsequently, the features are extracted by performing max pooling, and quantization is executed on continuous coordinates, represented as (x=16) within an open interval (excluding the endpoint, not a closed interval), where 'x' signifies the feature map stride and is rounded. However, this quantization process can potentially introduce conflicts when attempting to select the correct pixel-precise mask. To address this issue, RoIAlign is introduced to mitigate the harsh quantization inherent in RoIPool. The proposed alteration for RoI boundaries suggests using x=16, without enclosing it in brackets [x=16]. Bilinear interpolation will then be employed to

sample each bin location, and instead of selecting the maximum input, the average value will be computed.

2.3. Network Design:

In order to refine the mask approach, several design options were explored:

a) Conducting Feature Extraction Across the Entire Image with a Single Mask Prediction: This approach involves extracting features from the entire image and subsequently making a single mask prediction.

b) Creating a Localized RoIPool for Classification and Regression, and Then Applying Separate Mask Predictions to Each Segment: This layout entails generating a localized Region of Interest Pooling (RoIPool) for classification and regression. Following this, distinct mask predictions are applied to each segment.

An evaluation was carried out on ResNets and ResNeXt networks with depths of either 50 or 101 layers. Initially, the implementation was executed on Faster RCNN, using features extracted from both the final and the 4th convolutional layer (also referred to as the C-4 stage) of ResNets. This led to the development of ResNet-50, also known as ResNet-50-C4.

Lin et al. [18] introduced an alternative network known as the Feature Pyramid Network (FPN) [18]. FPN employed an up-down design with close connections to construct an in-network feature pyramid derived from a single-scale input. In Faster R-CNN with FPN support, RoI features are obtained from distinct levels of the feature pyramid based on their scale. However, the remaining architecture closely resembles that of the vanilla ResNet [19].

Integrating a ResNet-FPN backbone for feature abstraction with Mask RCNN yields notable improvements in both accuracy and processing speed. The specific details of the Mask architecture are illustrated in Figure 4. The head on the ResNet-C4 backbone encompasses the 5th stage of ResNet, referred to as the 9-layer 'res5'. Generally, an increase in the number of stages leads to greater accuracy.

Table 1: Comparison of various algorithms

Algorithm	Observations
Mask RCNN [10]	i. This algorithm stands out as the most versatile and straightforward in its execution among those discussed thus far.

	ii. It represents a notable advancement over previous algorithms like fast RCNN, Faster RCNN, and OCR. iii. Demonstrating impressive performance even with images captured at 5 frames per second, it holds the potential to revolutionize tasks such as training machines to detect human poses. iv. Recognized as the victor in the COCO 2016 datasets, it is distinguished as the swiftest algorithm in its class.
Fast RCNN [15]	i. Due to its incompatibility with humanoid robots, an enhanced version of HOG (Histogram of Oriented Gradients) was devised, resulting in the development of the fast RCNN along with a novel Region Proposal Network algorithm. ii. A proficient detection system targeting humans was effectively constructed by combining the capabilities of Fast RCNN with the VGG network. iii. The Fast RCNN approach yielded an impressive recognition rate of 97.3%, signifying a substantial 7% enhancement over the performance of the HOG algorithm. Additionally, the miss rate saw a notable reduction of 4%.
Haar-like feature [20] [21] [22] [23]	i. The detection of faces was specifically achieved through the utilization of Haar-like features. ii. This distinctive feature excels in capturing the co-occurrence of features within the image. iii. The facial recognition algorithm incorporating Haar-like features exhibited a remarkable 27% reduction in errors compared to those

	algorithms that did not employ this technique.
Faster RCNN [24]	i. Faster RCNN demonstrated remarkable performance, achieving an impressive 91% accuracy, which was a notable accomplishment in itself when compared to CNNs. ii. In instances where goals were not aligned, RCNN exhibited some discrepancies in shared features during the classification within a local cluster. iii. Faster RCNNs facilitated the implementation of multitasking learning. iv. To enhance the performance of the hard-false classifier, the RCNN was integrated with the Decoupled Classification Refinement (DCR) network. v. An extension was applied to the Faster RCNN network to encompass generic object detection, spanning face detection, concatenation, multi-scale training, hard negative mining, and precise configuration of anchor sizes for the Region Proposal Network (RPN). This research concluded that the FDDB test was the most fitting evaluation metric for the face detection algorithm.
OCR [25] [26] [27]	i. This algorithm was purpose-built with a focus on character identification. ii. At present, the algorithm is optimized for recognizing printed text. However, there is room for potential expansion to include handwritten notes in its scope. iii. Its execution hinges on the utilization of a Fuzzy Logic controller. iv. The algorithm demonstrated a remarkably low error rate, paving the

	way for accurate text detection within real-world images.

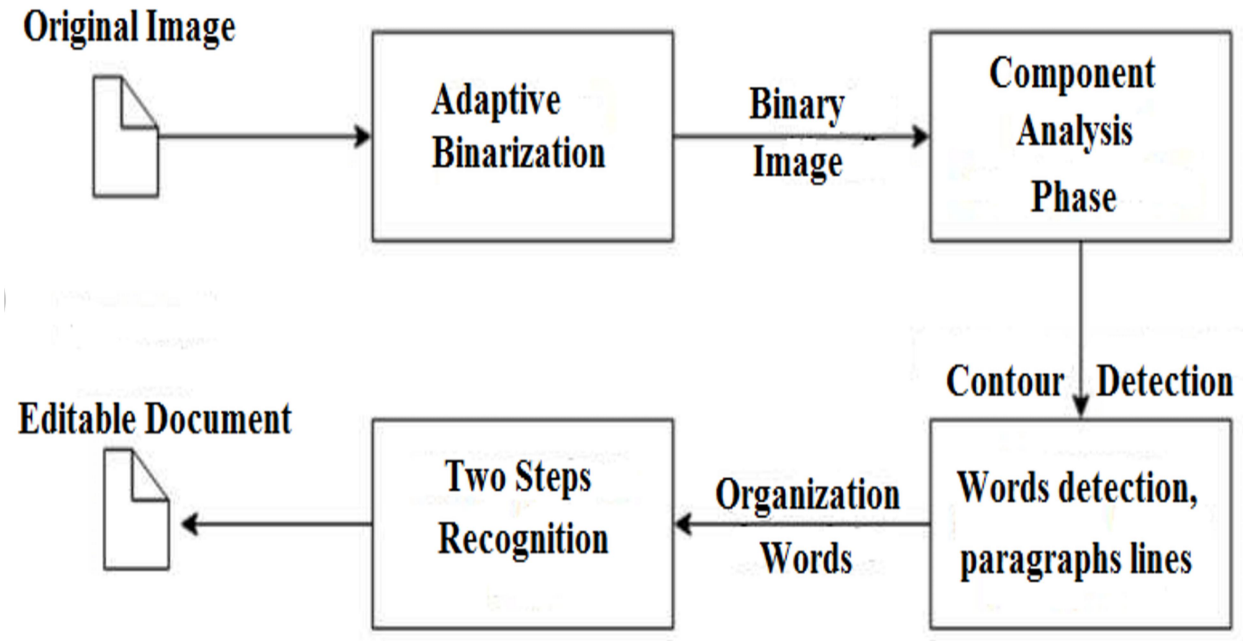

Figure 2: OCR process [25]

3. FUTURE SCOPE

Strides have been taken in the realm of object detection thanks to the efforts of Facebook AI researchers, employing the Mask RCNN algorithm which has achieved an impressive 86% accuracy in precise object identification. The potential of this object detection algorithm can further be harnessed for text detection within images or scenes. Moreover, its application can be extended to live monitoring systems like CCTVs and hawk-eye systems, ensuring swifter object recognition. Real-time object identification holds substantial promise for practical applications, such as resolving conflicts in sports like determining fouls or identifying the winning cyclist in a race. While significant progress has been made in detecting moving objects at 5 frames per second (fps), this speed may be considered sluggish for real-world scenarios. Therefore, there arises a necessity for an algorithm capable of even faster object identification.

4.CONCLUSION

The discussion highlights the pivotal role of computers in addressing a wide array of challenges, spanning from robotics and healthcare, encompassing the identification of tumors and heart diseases, to the practical applications of image segmentation and processing within the realm of facial recognition. It is evident that Mask RCNN stands as a burgeoning field, successfully living up to its name. However, with the advent of the enhanced Faster RCNNs, we witness expedited image detections that exhibit heightened accuracy compared to Mask RCNNs. Nonetheless, it's crucial to acknowledge that Mask RCNN remains the foundational framework for all these advancements. There is an optimistic outlook for further progress in the realm of object detection and text segmentation.

5. REFRENCES

[1] K. Suzuki, H. Abe, H. MacMahon, and K. Doi, “Image-processing technique for suppressing ribs in chest radiographs by means of massive training artificial neural network (MTANN),” IEEE Trans. Med. Imaging, vol. 25, no. 4, pp. 406–416, 2006.

[2] M. W. Eysenck and M. T. Keane, “Chapter 3: Object and Face Recognition,” Cogn. Psychol. A Student’s Handb., pp. 79–118, 2010.

[3] I. Arel, D. Rose, and T. Karnowski, “Deep machine learning-A new frontier in artificial intelligence research,” IEEE Comput. Intell. Mag., vol. 5, no. 4, pp. 13–18, 2010.

[4] “Let’s Dive in the World of Machine Learning – Yudiz Solutions – Medium,” 2019.

[5] A. E. Pandelea, M. Budescu, and G. Covatariu, “Image Processing Using Artificial Neural Networks,” Bul. Institutului Politeh. Din Iaşi, vol. 61, no. Lxv, pp. 10–21, 2015.

[6] C. A. L. Bailer-Jones, R. Gupta, and H. P. Singh, “An introduction to artificial neural networks,” 2001.

[7] M. Liao, B. Shi, X. Bai, X. Wang, and W. Liu, “TextBoxes: A Fast Text Detector with a Single Deep Neural Network,” 2016.

[8] J. W. Johnson, “Adapting Mask-RCNN for Automatic Nucleus Segmentation,” pp. 1–7, 2018.

[9] S. Mehri et al., “SampleRNN: An Unconditional End-to-End Neural Audio Generation Model,” pp. 1–11, 2016.

[10] P. Ammirato and A. C. Berg, “A Mask-RCNN Baseline for Probabilistic Object Detection,” CVPR Work., 2019.

[11] X. Wang, A. Shrivastava, and A. Gupta, “A-Fast- RCNN: Hard positive generation via adversary for object detection,” Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017, vol. 2017- January, pp. 3039–3048, 2017.

[12] K. He, G. Gkioxari, P. Dollar, and R. Girshick, “Mask R-CNN,” Proc. IEEE Int. Conf. Comput. Vis., vol. 2017-Octob, pp. 2980–2988, 2017.

[13] Z. Huang, Z. Zhong, L. Sun, and Q. Huo, “Mask R-CNN with pyramid attention network for scene text detection,” Proc. - 2019 IEEE Winter Conf. Appl. Comput. Vision, WACV 2019, pp. 764–772, 2019.

[14] T. Y. Lin et al., “Microsoft COCO: Common objects in context,” Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 8693 LNCS, no. PART 5, pp. 740–755, 2014.

[15] R. Girshick, “Fast R-CNN,” Proc. IEEE Int. Conf. Comput. Vis., vol. 2015 Inter, pp. 1440–1448, 2015.

[16] S. S. Kumar, P. Rajendran, P. Prabaharan, and K. P. Soman, “Text/Image Region Separation for Document Layout Detection of Old Document Images Using Non-linear Diffusion and Level Set,” Procedia Comput. Sci., vol. 93, no. September, pp. 469–477, 2016.

[17] J. Liu, X. Liu, J. Sheng, D. Liang, X. Li, and Q. Liu, “Pyramid Mask Text Detector,” 2019.

[18] T. Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature pyramid networks for object detection,” Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017, vol. 2017- January, pp. 936–944, 2017.

[19] L. (Stanford U. Sun, “ResNet on Tiny ImageNet,” pp. 1--7, 2012.

[20] Q. Chen, N. D. Georganas, E. M. Petriu, K. Edward, A. Ottawa, and C. Kin, "111Haar_Feature.Pdf," 2007.

[21] T. Hoang Ngan Le, Y. Zheng, C. Zhu, K. Luu, and M. Savvides, "Multiple Scale Faster-RCNN Approach to Driver's Cell-Phone Usage and Hands on Steering Wheel DetectionHoang Ngan Le, T., Zheng, Y., Zhu, C., Luu, K., & Savvides, M. (2016). Multiple Scale Faster-RCNN Approach to Driver's Cell-Phone Usage and Hands on Stee," Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Work., pp. 46–53, 2016.

[22] S. Zhang, C. Bauckhage, and A. B. Cremers, "Informed haar-like features improve pedestrian detection," Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., pp. 947–954, 2014.

[23] T. Mita, T. Kaneko, and O. Hori, "Joint Haar-like features for face detection," Proc. IEEE Int. Conf. Comput. Vis., vol. II, pp. 1619–1626, 2005.

[24] X. Sun, P. Wu, and S. C. H. Hoi, "Face detection using deep learning: An improved faster RCNN approach," Neurocomputing, vol. 299, pp. 42–50, 2018.

[25] L. Converso and S. Hocek, "Optical character recognition," J. Vis. Impair. Blind., vol. 84, no. 10, pp. 507–509, 1990.

[26] R. Singh, C. S. Yadav, P. Verma, and V. Yadav, "Optical Character Recognition (OCR) for Printed Devnagari Script Using Artificial Neural Network," Int. J. Comput. Sci. Commun., vol. 1, no. 1, pp. 91–95, 2010.

[27] J. Mao, "Case study on Bagging, Boosting, and Basic ensembles of neural networks for OCR," IEEE Int. Conf. Neural Networks - Conf. Proc., vol. 3, pp. 1828–1833, 1998.