Comparative Study of Different Techniques for Text as well as Object Detection from Real Time Images

Monika Kapoor[1], Er. Saurabh Sharma[2]

[1]Student, Dept of computer science Engineering, Sri Sai University Palampur (H.P) India
[2]Assitant Professor, Dept. of Computer science and Engineering, Sri Sai University Palampur (H.P) India

Abstract - Since the introduction of artificial intelligence, it has been a major curiosity of the developers to make computers think more like humans. Artificial Neural Networks was developed with the view in mind to make computers to do so. These intelligent machines can be used in the field of robotics, medicines, industry. Since the introduction of image recognition algorithms by face-book AI researchers image identification and recognition has become a curios field among the developers. The main objective of the image identification software(s) is to differentiate the components of the image or the scene and tell them apart. Many algorithms such as OCR, RCNN, Mask RCNN, Fast RCNN, Faster RCNN algorithms were developed to identify and classify images into the categories as desired by the programmer. Here in this paper, we would throw some light on the aforementioned different types of Image Processing algorithms and try to determine which of these are best suited. Researchers have proved that Mask RCNN, which is a better version of Faster RCNN proved to the best suited algorithm in the field of object detection in the real time.

***Keywords**: Artificial Neural Networks, Machine Learning, Object Identification, Scene Identification, OCR, Mask RCNN, RCNN.*

1. INTRODUCTION

Realworld applications are looking forward to make use of computers to make tasks easier for people in real world purposes. Due to its easiness in the usage computers are used in wide variety of fields such as robotics, medicine and advanced computing. Likewise, if robots are feed data about their surroundings then they can be better informed about their nearby environments to handle the situations. In a similarly way, if a computer can been taught how to differentiate and recognize tumor cells then it can be used for the identification of the tumors and cancer cells from the hundreds of pictures and can reduce the labor of manually identifying the tumors[1]. Likewise, in the case of the self-driven cars it becomes important for the system to identify the objects not only in the still images but in real time and take measures accordingly. Object detection aims to learn the concept of visual models in an image[2]. The ability to model the various deformations, inclusions, and other class variations while handling the large amount of data simultaneously under several conditions, is the main objective through this type of study.

Machine Learning (ML) [3] is a branch of artificial intelligence that is targeted towards the training of machines; designing the algorithms to handle robots etc. these machines learn to operate on their own after training on datasets. Machines are taught to take data driven decisions through self-calculation and self-observation instead of being exclusively programmed for a special task they are programmed for performing a number of tasks simultaneously. The a self-iterating algorithm is developed in such a way that it learns and improves itself. If a new input data is encountered by the ML algorithm, it makes a forecast based on the model. This forecast is validated for precession and if the precision is within acceptable choice, the Machine Learning algorithm is deployed. If the precision value is not accepted, the Machine Learning algorithm is trained to perform a number of iterations couple of times with an enlarged training data set.

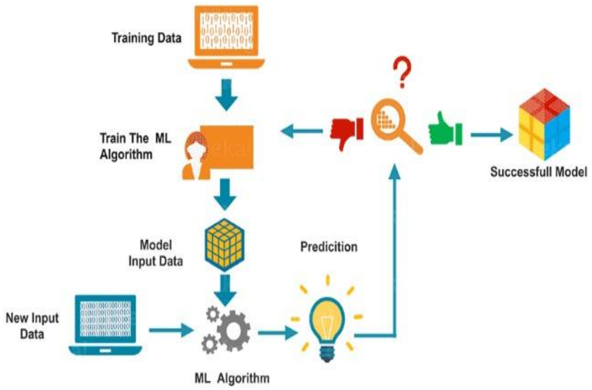

Figure 1: Machine learning algorithm [4]

The question arises as how to train the machines so that they can be made to think like humans. The answer to this question is provided by artificial neural networks (ANNs).

ANNs [5] are very helpful tools that have a vast use in artificial intelligence particularly in the field where facial recognitions are to be used and while doing so they reduce the errors due to human interference. As the name suggests ANNs are efficient computing systems whose concepts are similar to that of a parallel distributed processing system, and connectionist systems. ANNs take a large set of data altogether and this data is then passed through a set of interlinked units commonly known as nodes or neurons to display the desired output.

Neural networks [6] on the other hand are also perfectly apt to help people solve complicated problems in real-life situations. They are capable of learning and developing the co-relations among inputs and outputs that are nonlinear and complicated; generalize and infer; show hidden co-relations, patterns and predictions; and generate highly volatile data (including those of the financial time series) and variances that are needed to predict extremely rare events (like that of fraud detection). Neurons are joined to other neurons through a special link known as connection link.

These neurons are most commonly used in the field of facial recognition software, hence comes a new branch of image processing. Image Processing handles management of images through a computer. Just like the computer deals with the data inputs of numbers etc. in this case the computers are made to deal with images instead of a raw data. It aims on developing a computer algorithm that performs processing on images. The input will be a digital image and the program process that picture using written algorithms, and gives an output as picture itself.

1.1BACKGROUND

A more real time approach in the direction of the advancement of artificial neural networks was given by PDP group. It showed neural-based classification system, an approach to perform neutral facial image recognition using Parallel Hopfield Neural Networks. Alexandrina-Elena Pandelea [5] et.al, in 2015 used image processing in neural networks for geotechnical engineering, landslides problem was resolved by training network with ASTER images and GIS and a generation after learning maps. Minghui Liao [7] et.al 2016 presented the text box approach for faster scene detection. It is an end to end fast scene detection model that uses high accuracy and precision in a single network without any after processing on the image except for a standard non maximum suppression this method was fast for detecting the text at about 0.09 seconds that was considered to a highly fast text detection during its time. The Mask-RCNN model was developed in 2017 an extension to the Faster-RCNN model for semantic Division, object localization, and object instance Division of natural images. Mask-RCNN outperformed every single one of the pre-existing models in every way in the 2016 COCO Challenge [8], a large-scale object detection, segmentation, and captioning challenge. Kaiming He [9] et.al in the year 2018 presented a relatively simple, stretchy, and easier outlines for object case segmentation. Their technique very accurately detects the objects in the picture along with generation of high-quality segregation of the mask for every occurrence. Results were charted for instance segmentation, bounding box object detection, and person key point detection in the COCO challenge. Without any hesitation, Mask R-CNN outperformed all existing, single-model entries on every task, including the COCO 2016 challenge winners. The HOG and SIFT [10] pyramids have been used in numerous works for instance classification, object uncovering, human pose approximation, and more. Doll´ar [11] et al. established fast pyramid calculation by first computing a meagerly sampled (in scale) pyramid and then incorporating absent planes. Before HOG and SIFT, initial work on appearance recognition with ConvNets calculated narrow systems over image pyramids to detect appearances across scales. Some recent works include ResNet, Inception-ResNet and ResNetXt for object detection.

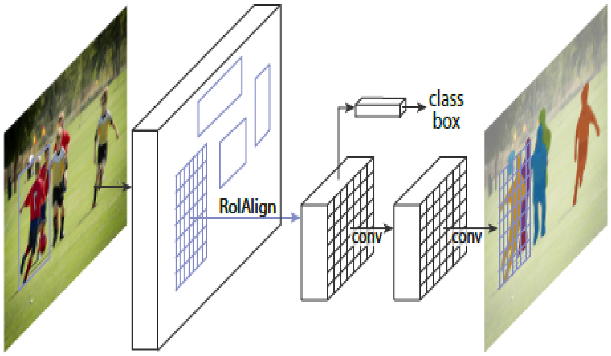

Figure 2: The Mask R-CNN framework for instance segmentation [12]

1.1.1Mask RCNN

Mask RCNN is the most advanced deep-learning algorithm for object detection and instance segmentation that is gaining popularity rapidly and is main focus of our review. Mask RCNN [13] achieved excellent results on the MS-COCO dataset [14]. Mask-RCNN is a double-stage image recognition process. Firstly, features are grasped from the image using a backbone Convolutional neural network (CNN), and class agnostic regional ideas are predicted. These ideas are then filtered and grouped in the second stage, to become either labeled bounding boxes for object detection or segmentation masks for the example segmentation task.

It can separate different objects in an image or a video. In large part, these developments have been motivated by powerful pre-cursor systems, i.e. the Fast/Faster RCNN [11] and Fully Convolutional Network (FCN) [15] outlines for object detection and semantic segmentation, respectively. The aim is to bring up a comparably allowing outline for instance segmentation. Instance segmentation [16] is stimulating because it requires the precise finding of all items in an image while also precisely segmenting each occurrence. In addition, single-shot models such as YOLO and SSD have enabled object detection to befall at hustles up to 100-1000 times quicker than area proposal-based algorithms.

Mask R-CNN [12] known as the object mask after objects class label and the bounding box is added in the mask RCNN makes it the better version of Fast RCNN. M-RCNN requires the mining of very finer pixels in an object. Furthermore, Mask RCNN focuses more on the pixel to pixel arrangement in the layout of an object that was the missing piece of the puzzle in both fast and faster RCNN. Faster R-CNN introduces us to the concept of the two stage detection systems one being Region Proposal Network (RPN) and another one called the RoIPool (Region of Interest).

The whole procedure of object identification through Mask RCNN can be simplified into the following: initially saying that Mask RCNN follows the same procedure as Faster RCNN i.e. division of the image into RPN followed by the second step successively by predicting the class and box offset, in addition to this mask RCNN also gives a binary mask for each RoI. This helps in the mask predictions. Each RoI is sampled by the logic as $L = L_{cls} + L_{box} + L_{mask}$. Where bounding box losses are represented by L_{cls}, bounding box region is given by L_{box}, and the new introduced mask feature is incorporated in the term L_{mask}. The mask part of the formula is a dimensional matrix of K_m^2 for each of the RoI. The K_m^2 is a binary output of the m x m matrix one for each class. To this further refinement will be done and losses are cut in the name of average binary cross entropy loss this will remove the overlapping of the boxes. The RoI associated to the ground truth class k, L_{mask} is only defined on the K^{th} mask disregarding the other losses. L_{mask} will generate the classes for every segment without interfering with the other class this is important as there is a huge reliability on the classes for the masked prediction. It is important here to mention that this step separate the Mask RCNN with the FCNs thus avoids the cross-entropy loss. These statements are well supported by the experimentations done by researchers.

Mask Representation: instead of the class labels and the bounding boxes here the mask is collapsed into short output vectors of fully connected layers this this mask is spatially structured and pixel to pixel connected. A prediction is made of m x m mask is made without damaging the dimensional vectors in the spatial arrangement. RoI feature requires the pixel to pixel corresponding link to make a better mask prediction.

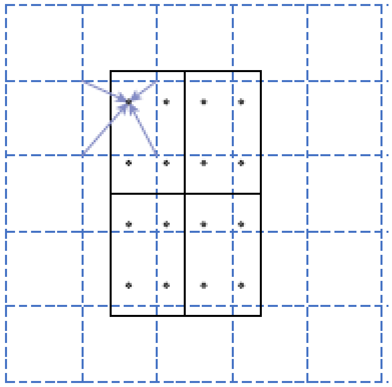

Figure 3: RoIAlign: the feature map is represented by the dashed grid whereas the region of interest (RoI) is represented by the solid grid and the dots represent the sampling points.[17]

1.1.2RoIAlign

RoIPool is a pretty basic and straight forward feature map like 7x7 or 9x9. Like calculating the maximum this RoI pool will firstly quantize the whole spatial arrangement into the distinct granules then this granule will be further divided into bins in space. So, we have very small spatially divided plane. Now by max pooling the features are attracted and quantization is performed on a continuous coordinate computing like (x=16) {open interval and not the closed interval} where is the feature map stride and is rounding. Due to this quantization there may give rise to the conflict in choosing the correct pixel accurate mask, but this problem may be addressed by a RoIAlign that will remove the harsh quantization of the pool. The change that is proposed is simple for RoI boundaries for the quantization use x=16 and not [x=16]. Bilinear interpolation will be used to sample each bin location and instead of the maximum input the average will be taken.

Network layout: to generalize the mask approach many layouts can be tested like a) feature extraction over an entire image and applying the mask prediction once or b) Creation of the localized RoIPool for classification and regression and then applying the mask predictions separately to each segment. Evaluation of ResNets and ResNeXt networks for the depth of 50 or 101 layes was done. Originally the implementation was done on Faster R-CNN with ResNets extracted features from the final and the 4th convolutional layer (also termend as C-4 stage). This is how ResNet-50 came into being; for example, ResNet-50-C4 is another name to the network. Another network was used by Lin et.al [18] called Feature Pyramid Network (FPN) [18]. FPN was based on up-down design with adjacent contacts to build an in-network feature pyramid from a single-scale contribution. Faster R-CNN with an FPN support takes RoI features from distinct organizations of the feature pyramid according to their scale, but otherwise the rest of the architecture is similar to vanilla ResNet [19]. Using a ResNet-FPN backbone for feature abstraction with Mask RCNN gives outstanding bonus in both correctness and speediness. Details of the Mask architecture are given in figure 4. The head on the ResNet-C4 backbone includes the 5-th stage of ResNet (namely, the 9-layer 'res5'). More will be the number of stages the greater the accuracy will be.

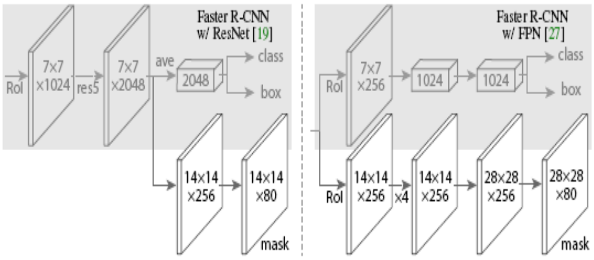

Figure 4: Head Architecture [12]

Table 1: Comparison of various algorithms

S No.	Algorithm	Remarks
1.	Haar-like feature [20] [21] [22] [23]	1. Haar-like feature was used for face detection specifically. 2. This feature captures the Feature Co-occurrence within the image. 3. Facial recognition algorithm that used Haar-like Feature was 27% less error free as compared to the others that did not use it.
2.	Fast RCNN [15]	1. Since the extension of HOG (Histogram of Oriented Gradients) is not suitable for the humanoid robots therefore fast RCNN along with new region Proposal Network algorithm was cooked up. 2. Human targeted detection system was successfully built based on the Fast RCNN along with VGG network was brought to light. 3. Fast RCNN gave 97.3% recognition rate, i.e. at least 7% improvement on HOG algorithm miss rate was also brought down by 4%.
3.	Faster RCNN [24]	1. Faster RCNN gave an outstanding result over the CNNs an astounding 91% accuracy which was in itself very remarkable. 2. When the goals were mismatched then the RCNN showed some errors in the features that were shared during the classification in a local cluster. 3. Multi-tasking learnings were made possible in the Faster RCNNs. 4. Decoupled Classification Refinement (DCR) network was used with the RCNN to improve the hard-false classifier. 5. An extension was made to the faster RCNN network for generic object detection including face detection, concatenation, multi-scale training, hard negative mining, and proper configuration of anchor sizes for RPN. It was concluded in this research that FDDB test is the most suitable for the face detection algorithm.

4.	OCR [25] [26] [27]	1. This algorithm was Tailor made for the identification of the characters. 2. For the time this algorithm is limited to the texts that are printed but an extension may be seen when it can be used for the hand written notes also. 3. This algorithm uses Fuzzy Logic controller for its execution. 4. The Error rate was minimum, which paved the way towards the text detection of the images in the real-world application.
5.	Mask RCNN [10]	1. It is the most generalized and simple in execution algorithm discussed so far. 2. It can be seen as an improvement on the fast RCNN, Faster RCNN, OCR, etc. 3. With the images at 5fps it can create wonders. It can be used to train the machines detect the human poses too. 4. winner of the COCO 2016 datasets it is labelled to the fastest algorithm

2. CONCLUSION

From the discussion it can be seen that computers help us solve our problems ranging from the field of robotics to that of the medicinal sector from identifying tumors and heart diseases to that of the real-world applications of image segmentation and image processing in the facial recognition sector. This will be safe to say that Mask RCNN is a developing branch that is able to live up to its name but now with the emergence of the new and improved faster RCNNs we can see faster image detections that can be accurate more as compared to the mask RCNNs. But Mask RCNN will always be the base of all these researches. And we can hope to see much improvement in the field of object detection and even text segmentation.

3. FUTURE SCOPE

Steps in the direction of object detection have been made by the Facebook AI researchers with the help of Mask RCNN algorithms that has reached an accuracy of 86% in identifying the objects correctly. This object detection algorithm can be extended for the text detection in the images/scenes. Also, object detection can be applied to the live CCTVs hawk-eye etc. systems for faster object identification. The object identification in real time at the phase is supposed to be helpful in practical life, take for instance, to resolve the conflicts, like that of fouls, or the winning cyclist in the race. The most helpful research involving moving objects was done at 5fps which is rather slow for the real-world applications. At this speed it is not possible for the real time scene detection algorithm to work. So, there is a need for an algorithm that can identify the objects at a much faster rate.

4. ACKNOWLEDGMENT

This review was made possible under the kind guidance of Er. Saurabh Sharma, Asst Professor in the Department of Computer Science at Sri Sai University Palampur. The kind wisdom imparted by Guide helped me a lot in carrying out the research work my field. This review paper is a mere reflection of a small fragment of the research work. Lastly, I would like to

thank the Department of Computer Science at Sri Sai University Palampur for giving me their support and humble comments in the research work.

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.