International Journal of Research and Development in Applied Science and Engineering (IJRDASE)
ISSN: 2454-6844

A Review of Object Detection based on Artificial Intelligence Technique

Km. Deepu Yadav, Shyam Shankar Dwivedi
Computer science and Engineering Department,
Rameshwaram Institute of Technology & Management, Lucknow
kmdeepu621@gmail.com

Abstract: Object detection, as of one the most fundamental and challenging problems in computer vision, has received great attention in recent years. Its development in the past two decades can be regarded as an epitome of computer vision history. If we think of today's object detection as a technical aesthetics under the power of deep learning, then turning back the clock 20 years we would witness the wisdom of cold weapon era. This paper extensively reviews 400+ papers of object detection in the light of its technical evolution, spanning over a quarter-century's time. A number of topics have been covered in this paper, including the milestone detectors in history, detection datasets, metrics, fundamental building blocks of the detection system, speed up techniques, and the recent state of the art detection methods. This paper also reviews some important detection applications, such as pedestrian detection, face detection, text detection, etc, and makes an in-deep analysis of their challenges as well as technical improvements in recent years.

Keywords: Object detection, Computer vision, Deep learning, Convolutional neural networks, Technical evolution.

1. Introduction:

Object detection is an important computer vision task that deals with detecting instances of visual objects of a certain class (such as humans, animals, or cars) in digital images. The objective of object detection is to develop computational models and techniques that provide one of the most basic pieces of information needed by computer vision applications: What objects are where? As one of the fundamental problems of computer vision, object detection forms the basis of many other computer vision tasks, such as instance segmentation [1–4], image captioning [5–7], object tracking [8], etc. From the application point of view, object detection can be grouped into two research topics "general object detection" and "detection applications", where the former one aims to explore the methods of detecting different types of objects under a unified framework to simulate the human vision and cognition, and the later one refers to the detection under specific application scenarios, such as pedestrian detection, face detection, text detection, etc. In recent years, the rapid development of deep learning techniques [9] has brought new blood into object detection, leading to remarkable breakthroughs and pushing it forward to a research hot-spot with unprecedented attention. Object detection has now been widely used in many real-world applications, such as autonomous driving, robot vision, video surveillance, etc. Fig. 1 shows the growing number of publications that are associated with "object detection" over the past two decades.

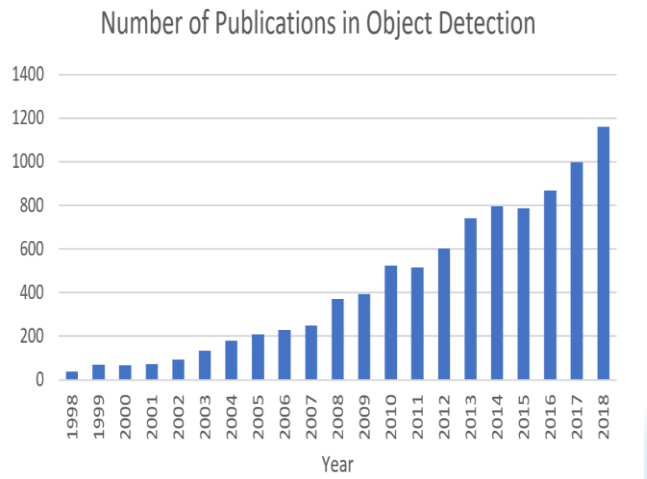

Fig 1. The increasing number of publications in object detection.

2. Literature Review:

Histogram of Oriented Gradients (HOG) feature descriptor was originally proposed in 2005 by N. Dalal and B. Triggs [12]. HOG can be considered as an important improvement of the scale-invariant feature transform [33, 34] and shape contexts [35] of its time. To balance the feature invariance (including translation, scale, illumination, etc) and the nonlinearity (on discriminating different objects categories), the HOG descriptor is designed to be computed on a dense grid of uniformly spaced cells and use overlapping local contrast normalization (on "blocks") for improving accuracy. Although HOG can be used to detect a variety of object classes, it was motivated primarily by the problem of pedestrian detection. To detect objects of different sizes, the HOG detector rescales the input image for multiple times while keeping the size of a detection window unchanged. The HOG detector has long been an important foundation of many object detectors [13, 14, 36] and a large variety of computer vision applications for many years.

The DPM follows the detection philosophy of "divide and conquer", where the training can be simply considered as the learning of a proper way of decomposing an object, and the inference can be considered as an ensemble of detections on different object parts. For example, the problem of detecting a "car" can be considered as the detection of its window, body, and wheels. This part of the work, a.k.a. "star-model", was completed by P. Felzenszwalb et al. [13]. Later on, R.

Girshick has further extended the star-model to the "mixture models" [14, 15, 37, 38] to deal with the objects in the real world under more significant variations. A typical DPM detector consists of a root-filter and a number of part-filters. Instead of manually specifying the configurations of the part filters (e.g., size and location), a weakly supervised learning method is developed in DPM where all configurations of part filters can be learned automatically as latent variables. R. Girshick has further formulated this process as a special case of Multi-Instance learning
[39], and some other important techniques such as "hard negative mining", "bounding box regression", and "context priming" are also applied for improving detection accuracy. To speed up the detection, Girshick developed a technique for "compiling" detection
models into a much faster one that implements a cascade architecture, which has achieved over 10 times acceleration without sacrificing any accuracy [14, 38].

As the performance of hand-crafted features became saturated, object detection has reached a plateau after 2010. R. Girshick says: "... progress has been slow during 2010-2012, with small gains obtained by building ensemble systems and employing minor variants of successful methods"[38]. In 2012, the world saw the rebirth of convolutional neural networks [40]. As a deep convolutional network is able to learn robust and high-level feature representations of an image, a natural question is whether we can bring it to
object detection? R. Girshick et al. took the lead to break the deadlocks in 2014 by proposing the Regions with CNN features (RCNN) for object detection [16]. Since then, object detection started to evolve at an unprecedented speed. In deep learning era, object detection can be grouped into two genres: "two-stage detection" and "one-stage detection", where the former frames the detection as a "coarseto- fine" process while the later frames it as to "complete in one step".

In 2014, K. He et al. proposed Spatial Pyramid Pooling Networks (SPPNet) [17]. Previous CNN models require a fixed-size input, e.g., a 224x224 image for AlexNet [40]. The main contribution of SPPNet is the introduction of a Spatial Pyramid Pooling (SPP) layer, which enables a CNN to generate a fixed-length representation regardless of the size of image/region of interest without rescaling it. When using SPPNet for object detection, the feature maps can be computed from the entire image only once, and then fixedlength representations of arbitrary regions can be generated for training the detectors, which avoids repeatedly computing the convolutional features. SPPNet is more than 20 times faster than R-CNN without sacrificing any detection accuracy (VOC07 mAP=59.2%).

Although SPPNet has effectively improved the detection speed, there are still some drawbacks: first, the training is still multi-stage, second, SPPNet only fine-tunes its fully connected layers while simply ignores all previous layers. Later in the next year, Fast RCNN [18] was proposed and solved these problems.

In 2015, R. Girshick proposed Fast RCNN detector [18], which is a further improvement of R-CNN and SPPNet [16, 17]. Fast RCNN enables us to simultaneously train a detector and a bounding box regressor under the same network configurations. On VOC07 dataset, Fast RCNN increased the mAP from 58.5% (RCNN) to 70.0% while with a detection speed over 200 times faster than R-CNN. Although Fast-RCNN successfully integrates the advantages of R-CNN and SPPNet, its detection speed is still limited by the proposal detection. Then, a question naturally arises: "can we generate object proposals with a CNN model?" Later, Faster R-CNN [19] has answered this question.

In despite of its high speed and simplicity, the one-stage detectors have trailed the accuracy of two-stage detectors for years. T.-Y. Lin et al. have discovered the reasons behind and proposed RetinaNet in 2017 [23]. They claimed that the extreme foreground-background class imbalance encountered during training of dense detectors is the central cause. To this end, a new loss function named "focal loss" has been introduced in RetinaNet by reshaping the standard cross entropy loss so that detector will put more focus on hard, misclassified examples during training. Focal Loss enables the one-stage detectors to achieve comparable accuracy of two-stage detectors while maintaining very high detection speed.

Building larger datasets with less bias is critical for developing advanced computer vision algorithms. In object detection, a number of well-known datasets and benchmarks have been released in the past 10 years, including the datasets of PASCAL VOC Challenges [10, 11] (e.g., VOC2007, VOC2012), ImageNet Large Scale Visual Recognition Challenge (e.g., ILSVRC2014) [12], MS-COCO Detection Challenge [53], etc. The statistics of these datasets are given some image examples of these datasets. Fig. 3 shows the improvements of detection accuracy on VOC07, VOC12 and MS-COCO datasets from 2008 to 2018.

The PASCAL Visual Object Classes (VOC) Challenges1 (from 2005 to 2012) [10, 11] was one of the most important competition in early computer vision community. There are multiple tasks in PASCAL VOC, including image classification, object detection, semantic segmentation and action detection. Two versions of Pascal-VOC are mostly used in object detection: VOC07 and VOC12, where the former consists of 5k tr. images + 12k annotated objects, and the latter consists of 11k tr. images + 27k annotated objects. 20 classes of objects that are common in life are annotated in these two datasets (Person: person; Animal: bird, cat, cow, dog, horse, sheep; Vehicle: aeroplane, bicycle, boat, bus, car, motor-bike, train; Indoor: bottle, chair, dining table, potted plant, sofa, tv/monitor). In recent years, as some larger datasets like ILSVRC and MS-COCO (to be introduced) has been released, the VOC has gradually fallen out of fashion and has now become a test-bed for most new detectors.

In the early time's detection community, there is no widely accepted evaluation criteria on detection performance. For example, in the early research of pedestrian detection [12], the "miss rate vs. false positives per-window (FPPW)" was usually used as a metric. However, the perwindow measurement (FPPW) can be flawed and fails to predict full image performance in certain cases [19].

In 2009, the Caltech pedestrian detection benchmark was created [19, 20] and since then, the evaluation metric has changed from per-window (FPPW) to false positives perimage (FPPI).

In recent years, the most frequently used evaluation for object detection is "Average Precision (AP)", which was originally introduced in VOC2007. AP is defined as the average detection precision under different recalls, and is usually evaluated in a category specific manner. To compare performance over all object categories, the mean AP (mAP) averaged over all object categories is usually used as the final metric of performance. To measure the object localization accuracy, the Intersection over Union (IoU) is used to check whether the IoU between the predicted box and the ground truth box is greater than a predefined threshold, say, 0.5. If yes, the object will be identified as "successfully detected", otherwise will be identified as "missed". The 0.5- IoU based mAP has then become the de facto metric for object detection problems for years.

After 2014, due to the popularity of MS-COCO datasets, researchers started to pay more attention to the accuracy of the bounding box location. Instead of using a fixed IoU threshold, MS-COCO AP is averaged over multiple IoU thresholds between 0.5 (coarse localization) and 0.95 (perfect localization). This change of the metric has encouraged more accurate object localization and may be of great importance for some real-world applications (e.g., imagine there is a robot arm trying to grasp a spanner). Recently, there are some further developments of the evaluation in the Open Images dataset, e.g., by considering the group-of boxes and the non-exhaustive image-level category hierarchies. Some researchers also have proposed some alternative metrics, e.g., "localization recall precision" [34]. Despite the recent changes, the VOC/COCO-based mAP is still the most frequently used evaluation metric for object detection.

Y. LeCun et al. have
made great contributions at that time. Due to limitations in computing resources, CNN models at the time were much smaller and shallower than those of today. Despite this, the computational efficiency was still considered as one of the tough nuts to crack in early times's CNN based detection models. Y. LeCun et al. have made a series of improvements like "shared-weight replicated neural network" [36] and "space displacement network" [37] to reduce the computationsby extending each layer of the convolutional network so as to cover the entire input image. In this way, the feature of any location of the entire image can be extracted by taking only one time of forward propagation of the network. This can be considered as the prototype of today's fully convolutional networks (FCN) [10, 11], which was proposed almost 20 years later. CNN also has been applied to other tasks such as face detection [12, 13] and hand tracking [14] of its time.

Early detection models like VJ detector and HOG detector were specifically designed to detect objects with a "fixed aspect ratio" (e.g., faces and upright pedestrians) by simply building the feature pyramid and sliding fixed size detection window on it. The detection of "various aspect ratios" was not considered at that time. To detect objects with a more complex appearance like those in PASCAL VOC, R. Girshick et al. began to seek better solutions outside the feature pyramid. The "mixture model" [15] was one of
the best solutions at that time, by training multiple models to detect objects with different aspect ratios. Apart from this, exemplar-based detection [36, 115] provided another solution by training individual models for every object instance (exemplar) of the training set.

As objects in the modern datasets (e.g., MS-COCO) become more diversified, the mixture model or exemplarbased methods inevitably lead to more miscellaneous detection models. A question then naturally arises: is there a unified multi-scale approach to detect objects of different aspect ratios? The introduction of "object proposals" (to be introduced) has answered this question.

In recent years, as the increase of GPU's computing power, the way people deal with multi-scale detection has become more and more straight forward and brute-force. The idea of using the deep regression to solve multi-scale problems is very simple, i.e., to directly predict the coordinates of a bounding box based on the deep learning features [20]. The advantage of this approach is that it is simple and easy to implement while the disadvantage is the localization may not be accurate enough especially for some small objects. "Multi-reference detection" (to be introduced) has latter solved this problem.

Most of the early detection methods such as VJ detector and HOG detector do not use BB regression, and usually directly consider the sliding window as the detection result. To obtain accurate locations of an object, researchers have no choice but to build very dense pyramid and slide the detector densely on each location.

The first time that BB regression was introduced to an object detection system was in DPM [15]. The BB regression at that time usually acted as a post-processing block, thus it is optional. As the goal in the PASCAL VOC is to predict single bounding box for each object, the simplest way for a DPM to generate final detection should be directly using its root filter locations. Later, R. Girshick et al. introduced a more complex way to predict a bounding box based on the complete configuration of an object hypothesis and
formulate this process as a linear least-squares regression problem [15]. This method yields noticeable improvements of the detection under PASCAL criteria.

Local context refers to the visual information in the area that surrounds the object to detect. It has long been acknowledged that local context helps improve object detection. At early 2000s, Sinha and Torralba [139] found that inclusion of local contextual regions such as the facial bounding contour substantially improves face detection performance. Dalal and Triggs also found that incorporating a small amount of background information improves the accuracy of pedestrian detection [12]. Recent deep learning based detectors can also be improved with local context by simply enlarging the networks' receptive field or the size of object proposals.

Global context exploits scene configuration as an additional source of information for object detection. For early time's object detectors, a common way of integrating global context is to integrate a statistical summary of the elements that comprise the scene, like Gist [10]. For modern deep learning based detectors, there are two methods to integrate global context. The first way is to take advantage of large receptive field (even larger than the input image) [20] or global pooling operation of a CNN feature. The second way is to think of the global context as a kind of sequential information and to learn it with the recurrent neural networks [14, 19].

3. Conclusion:

Remarkable achievements have been made in object detection over the past 20 years. This paper not only extensively reviews some milestone detectors (e.g. VJ detector, HOG detector, DPM, Faster-RCNN, YOLO, SSD, etc), key technologies, speed up methods, detection applications, datasets, and metrics in its 20 years of history, but also discusses the challenges currently met by the community, and how these detectors can be further extended and improved. Recent deep learning based detectors are becoming more and more sophisticated and heavily relies on experiences. A future direction is to reduce human intervention when designing the detection model (e.g., how to design the engine and how to set anchor boxes) by using neural architecture search. AutoML could be the future of object detection.

References:

[1] B. Hariharan, P. Arbel´aez, R. Girshick, and J. Malik, "Simultaneous detection and segmentation," in European Conference on Computer Vision. Springer, 2014, pp. 297–312.

[2] ——, "Hypercolumns for object segmentation and finegrained localization," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 447–456.

[3] J. Dai, K. He, and J. Sun, "Instance-aware semantic segmentation via multi-task network cascades," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3150–3158.

[4] K. He, G. Gkioxari, P. Doll´ar, and R. Girshick, "Mask rcnn," in Computer Vision (ICCV), 2017 IEEE International Conference on. IEEE, 2017, pp. 2980–2988.

[5] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in Proceedings of the IEEE conference on computer vision and pattern recog-nition, 2015, pp. 3128–3137.

[6] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in International conference on machine learning, 2015, pp. 2048–2057.

[7] Q.Wu, C. Shen, P.Wang, A. Dick, and A. van den Hengel, "Image captioning and visual question answering based on attributes and external knowledge," IEEE transactions on pattern analysis and machine intelligence, vol. 40, no. 6, pp. 1367–1381, 2018.

[8] K. Kang, H. Li, J. Yan, X. Zeng, B. Yang, T. Xiao, C. Zhang, Z. Wang, R. Wang, X. Wang et al., "T-cnn: Tubelets with convolutional neural networks for object detection from videos," IEEE Transactions on Circuits and Systems for Video Technology, vol. 28, no. 10, pp. 2896–2907, 2018.

[9] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," nature, vol. 521, no. 7553, p. 436, 2015.

[10] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on, vol. 1. IEEE, 2001, pp. I–I.

[11] P. Viola and M. J. Jones, "Robust real-time face detection," International journal of computer vision, vol. 57, no. 2, pp.
137–154, 2004.

[12] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Con- ference on, vol. 1. IEEE, 2005, pp. 886–893.

[13] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on. IEEE, 2008, pp. 1–8.

[14] P. F. Felzenszwalb, R. B. Girshick, and D. McAllester, "Cascade object detection with deformable part models," in Computer vision and pattern recognition (CVPR), 2010 IEEE conference on. IEEE, 2010, pp. 2241–2248.

[15] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," IEEE transactions on pattern analysis and machine intelligence, vol. 32, no. 9, pp. 1627–1645, 2010.

[16] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 580– 587.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in European conference on computer vision. Springer, 2014, pp. 346–361.

[18] R. Girshick, "Fast r-cnn," in Proceedings of the IEEE inter- national conference on computer vision, 2015, pp. 1440–1448.

[19] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in Advances in neural information processing systems, 2015, pp. 91–99.

[20] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 779–788.

[21] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in European conference on computer vision. Springer, 2016, pp. 21–37.

[22] T.-Y. Lin, P. Doll´ar, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie, "Feature pyramid networks for object detection." in CVPR, vol. 1, no. 2, 2017, p. 4.

[23] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Doll´ar, "Focal loss for dense object detection," IEEE transactions on pattern analysis and machine intelligence, 2018.

[24] L. Liu, W. Ouyang, X. Wang, P. Fieguth, J. Chen, X. Liu, and M. Pietik¨ainen, "Deep learning for generic object detection: A survey," arXiv preprint arXiv:1809.02165, 2018.

[25] S. Agarwal, J. O. D. Terrail, and F. Jurie, "Recent advances in object detection in the age of deep convolutional neural networks," arXiv preprint arXiv:1809.03193, 2018.

[26] A. Andreopoulos and J. K. Tsotsos, "50 years of object recognition: Directions forward," Computer vision and image understanding, vol. 117, no. 8, pp. 827–891, 2013.

[27] J. Huang, V. Rathod, C. Sun, M. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama et al.,

"Speed/accuracy trade-offs for modern convolutional object detectors," in IEEE CVPR, vol. 4, 2017.
[28] K. Grauman and B. Leibe, "Visual object recognition (synthesis lectures on artificial intelligence and machine learning)," Morgan & Claypool, 2011.
[29] C. P. Papageorgiou, M. Oren, and T. Poggio, "A general framework for object detection," in Computer vision, 1998. sixth international conference on. IEEE, 1998, pp. 555–562.
[30] C. Papageorgiou and T. Poggio, "A trainable system for object detection," International journal of computer vision, vol. 38, no. 1, pp. 15–33, 2000.
[31] A. Mohan, C. Papageorgiou, and T. Poggio, "Examplebased object detection in images by components," IEEE Transactions on Pattern Analysis & Machine Intelligence, no. 4, pp. 349–361, 2001.
[32] Y. Freund, R. Schapire, and N. Abe, "A short introduction to boosting," Journal-Japanese Society For Artificial Intelligence, vol. 14, no. 771-780, p. 1612, 1999.
[33] D. G. Lowe, "Object recognition from local scale-invariant features," in Computer vision, 1999. The proceedings of the seventh IEEE international conference on, vol. 2. Ieee, 1999,
pp. 1150–1157.
[34] ——, "Distinctive image features from scale-invariant keypoints," International journal of computer vision, vol. 60, no. 2, pp. 91–110, 2004.
[35] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," CALIFORNIA UNIV SAN DIEGO LA JOLLA DEPT OF COMPUTER SCIENCE AND ENGINEERING, Tech. Rep., 2002.
[36] T. Malisiewicz, A. Gupta, and A. A. Efros, "Ensemble of exemplar-svms for object detection and beyond," in Computer Vision (ICCV), 2011 IEEE International Conference on. IEEE, 2011, pp. 89–96.
[37] R. B. Girshick, P. F. Felzenszwalb, and D. A. Mcallester, "Object detection with grammar models," in Advances in Neural Information Processing Systems, 2011, pp. 442–450.
[38] R. B. Girshick, From rigid templates to grammars: Object detection with structured models. Citeseer, 2012.
[39] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning," in Advances in neural information processing systems, 2003, pp. 577–584.
[40] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Advances in neural information processing systems, 2012, pp. 1097–1105.