Deep Learning Algorithms-based Object Detection and Localization Revisited

Deep Learning Algorithms-based Object Detection and Localization Revisited

Safa Riyadh Waheed[1,2], Norhaida Mohd Suaib[1], Mohd Shafry Mohd Rahim[1,3], Myasar Mundher Adnan[4], and A. A. Salim[5]

[1]Big Data Center & Faculty of Engineering, School of Computing, Universiti Teknologi Malaysia, 81310, Johor Bahru, Malaysia
[2]Computer Techniques Engineering Department, Faculty of Information Technology, Imam Jaafar Al-Sadiq University, Baghdad, Iraq
[3]UTM-IRDA Digital Media Institute of Human-Centred Engineering, Universiti Teknologi Malaysia, 81310, Johor Bahru, Malaysia
[4]Islamic University, Najaf, Iraq
[5]Laser Center & Physics Department, Faculty of Science, Universiti Teknologi Malaysia, 81310, Johor Bahru, Malaysia

Email: safa_albdeary@hotmail.com

Abstract. The computer vision (CV) is an emerging area with sundry promises. This communication encompasses the past development, recent trends and future directions of the CV in the context of deep learning (DL) algorithms-based object detections and localizations techniques. To identify the object location inside an image and recognize it by a computer program as fast as the human brain the machine learning and DL techniques have been evolved. However, the main limitations of the machine are related to the prolonged time consumption to handle vast amount of data to perform the same task as the human brain. To overcome these shortcomings, the convolution neural networks (NNs)-based deep NN has been developed, which detects and classifies the object with high precision. To train the deep NNs, massive amount of data (in the form of images and videos) and time is needed, making the computational cost of the CV very high. Thus, transfer learning techniques have been proposed wherein a model trained on one task can be reused on another linked task, thereby producing excellent outcomes. In this spirit, diverse DL-based algorithms have been introduced to detect and classify the object. These algorithms include the region-based convolutional NN (R-CNN), fast R-CNN, Faster R-CNN, mask E-CNN and You Only Look Once. A comparative evaluation among these techniques has been made to reveal their merits and demerits in the CV.

1. Introduction

In the field of computer vision (CV), various algorithms and approaches are used to read, process, analyse and understand a given image. In addition, it generates a new kind of symbolic information in the form of high dimensional data that fits into the natural surroundings for further decision making [1]. Thus, computer scientists and engineers have been constantly motivated toward the CV studies where a machine can process the visual information remarkably similar to the human. Meanwhile, an image can

be understood in different formats (for example the symbolic or numeric) and may be obtained via varieties of the mathematical models including the statistical, geometrical, learning or combination of two or more of them [2]. The CV is an emergent area with immense applications potential in the domain of machine learning, medical imaging, industrial control systems, physics and so forth. It is divided into various subfields such as the event and object detection (OD) as well as tracking, object recognition, scene reconstruction, image restoration, video applications and motion estimation to cite a few [3].

One of the most vital mechanisms of the CV is the OD that is essential prior to the applications including the recognition, categorisation, event detection, or objects searching in a video stream for processing. Furthermore, the identification of an object of interest in a motion video stream or a set of images remains a major challenge due to its ever-growing demand in the smartphones and other devices applications. For the supplementary analysis and decision making, these recognitions are centred to the people, landmarks, buildings, or any specific object of interest [4]. In short, the recognition of the visual objects still remains a computationally intensive and challenging issue. This problem is due to the formation of infinite number of two-dimensional images on the retina for each object in the word that frequently alters with the directions, positions, lighting conditions and backgrounds concerning the viewer's perspectives [5]. These variations can be majorly ascribed to the deformation of the non-rigid visual characteristics. In addition, the intra-class alterations related to the shape are another attributes that intensify the problem [6]. The OD qualities (such as the detection rate, error rate, time and precision metrics) obtained by the CV are decided by the time and precision trade-off.

A comprehensive OD system has three essential characteristics such as the feature extraction, object recognition, and object localization. The performance of any OD system is somewhat influenced by each of these attributes. Most of the previous studies focused on a single feature of the OD system. The supplementary factors such as the change in scale and rotation that cause the data imbalance can further hamper the OD system performance. Besides, the performance of the OD system can be improved using a novel learning descriptor that is fast and unique with invariant features. The improvement of the ensemble-based learning enables it less sensitive to the data imbalance. In addition, any change of the detector for the original interest points can enhance the detection performance. Over the years, several studies have been performed to develop a precise OD scheme in the CV with improved performance, where the feature extraction, object recognition, and object localization were focused. Despite many dedicated efforts an all-inclusive OD system in the CV enclosing the three primary elements (features, recognition, and localization**)** remains deficient. It is now customary to discuss the evolution of these three essential components.

1.1 Features Extraction

Depending on the spatial information around the object and the extraction mechanisms, the features are classified into two categories wherein the spatial features can be local or global [4] [2]. The features can be acquired via the process of extraction, selection and composition [9]. Any feature is characterized by a unique attribute to describe a detectable object which is comprised of colors, edges, corners, and textures [10]. It addition, other information are needed to describe the entire image (global features) or a region of interest within the image (local features) the features play significant role in the OD and directly affect the quality of detection. According to feature classes for the ODs within the images, the objects must be characterized by the features that allow further classification [11].

1.2 Object Recognition

This step detects or identifies the specific object related to the image irrespective of its class [12] [10] wherein the object can be recognized via the learning or matching technique. In matching technique, the object is identified using the distance between the features concerning the target and stored template. Conversely, the notion of the learning technique (called machine learning) has been evolved from the capacity of human to classify or predict new instances or tasks using the past experience, knowledge and analytical observation. The machine learning techniques are further divided into two categories

depending on the types of learning or number of learners engaged. The number of learning classifiers can be single or multiple depending on the ensemble. These learning techniques include unsupervised, semi-supervised, supervised and reinforced as explained briefly hereunder.

The supervised learning is the teacher-based that depends on the knowledge of various object patterns belonging to a set of examples. This includes the storage of sufficient set of templates to converge the input into a desired output (teacher) by generating a proper function or rule. The output connected to the supervised learning can either be continuous (regression) or labelled class (classification) [10]. The supervised learning can further be classified into four broad categories such as the based on the logic, perceptron, statistics, instance and SVM [13]. In contrast, the unsupervised learning aims to identify the rule or structure in the unlabeled data (without target attributes) without any guidance for evaluating the desired solution (for example the error between the desired target and output). The semi-supervised learning unifies the unsupervised and supervised learning. This technique may not require all the labelled examples (target) where few of them are sufficient for the learning. The reinforced learning is an active mechanism that involves the interaction with the surrounding environment through a series of actions instead of labelled targets [14]. Figure 1 shows the categories and sub-categories of various machine learning techniques.

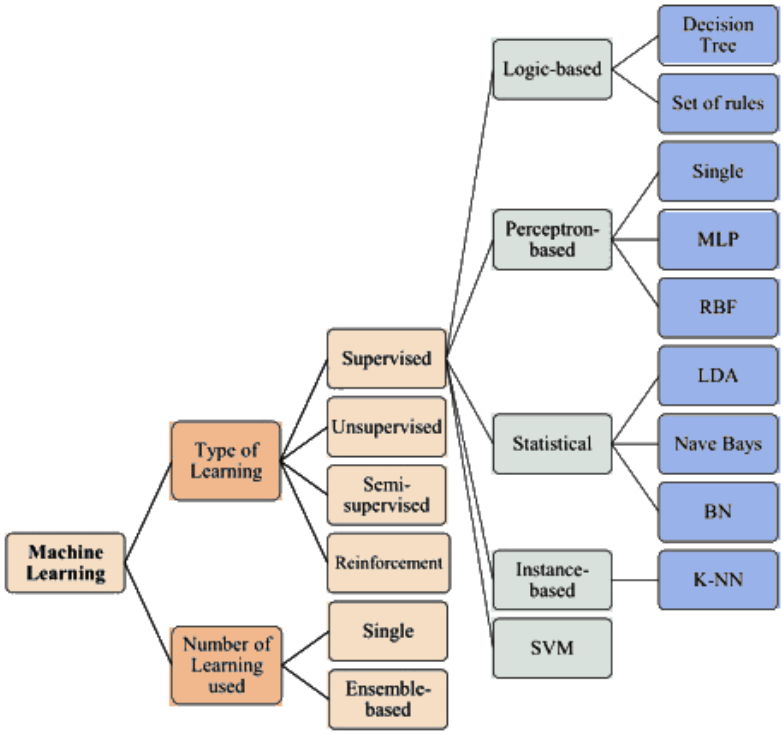

Figure 1: Classifications of the machine learning techniques [15, 16]

1.3 Object Localization

In this phase the objects are localized for accurate CV. In fact, the object classification and object localization are two different mechanisms. The former one determines the presence of the objects in the scene, whereas the later one (object localization) decides the locations of the objects [17] [18] [19]. Thus, the object localization is an intricate task towards the understanding and analyses of the high-level components of the automatic image in the CV [20] [21]. Yet again, the variations in the intra-class [22], scaling, viewpoint, poor image quality, background clutter and occlusion make this task more challenging [23]. Therefore, all the existing state-of-the-art methods for the object localization that

satisfy the requirements of both accuracy and speed remain inferior for the real-time applications [24] [25]. The objects can be localized via the center points, contours (Lampert et al., 2008) or bounding boxes [22]. The sliding-window is the most common technique used to localize the objects. Yet, this technique has many shortcomings:

i) Low image dimension (for example 320 × 240) containing a billion rectangular sub-images. Therefore, the sub-images are required for large image size that is computationally too expensive [19].
ii) It scans all possible scales and positions, making the technique computationally too costly without any assured localization [25].
iii) Due to the diverse shapes of the objects, no sub-window can narrowly localize the object of interest [26]. The bounding box method is also inaccurate to localize the objects [27].
iv) The number of windows can be decreased to accelerate the localization process, but at the expense of the accuracy.

1.4 Object Detection

The OD is a CV-based technology that enables to identify and locate the classes of the objects in an image (for instance faces, cars, people, and so on) [28-30]. However, the object recognition classifies an instance of an object into its class [12]. The objects can be detected by combining the CV and image processing techniques [31]. In addition, the machine learning may be used in an OD system. The OD is a vital process in various tasks that includes the object and events detections, objects recognition and tracking, video indexing, motion estimations, image registrations [32], image restorations [3], image retrievals, localization of the cameras, and reconstructions of the 3D scenes [34]. Based on these factors, the detection and discovery of the specific objects in the sequences of the images or videos became increasingly significant to numerous applications with the advent of mobile phones and other audio-visual technologies. As aforementioned, the objects detection in the CV is used to search persons, buildings, places and various other things of interest for further analyses and decision making [2, 4]. In this rationale, the OD have been exploited in a broad array of applied fields such as the industrial (machine inception and robotics control) [31], biometric security (face verification) [32], surveillance [33] and automotive safety.

The visual OD is a complicated and tedious computational problem due to the formation of limitless numbers of different 2-D images onto the human retina by any object. In addition, the image features change with the alteration in the object positions, poses, lighting, and backgrounds with respect to the viewer [3]. These alterations are mainly due to the non-rigid deformations, intra-class variability of the shapes and other visual properties [6]. Efforts have continually been made to resolve these computational issues. Presently, the OD issues are classified into 3 types including the geometric transformations, photometric transformations, and auxiliary disturbances [35, 36]. The geometric transformations involve the changes in the rotation and scaling [37, 38]. The photometric transformations deal with the alterations in the illumination. The background confusion [16] and the instances of such disturbances include the image blurring [40], occlusions [41] and intra-class variations [42]. Figure 2 illustrates the most common challenges in the image detection.

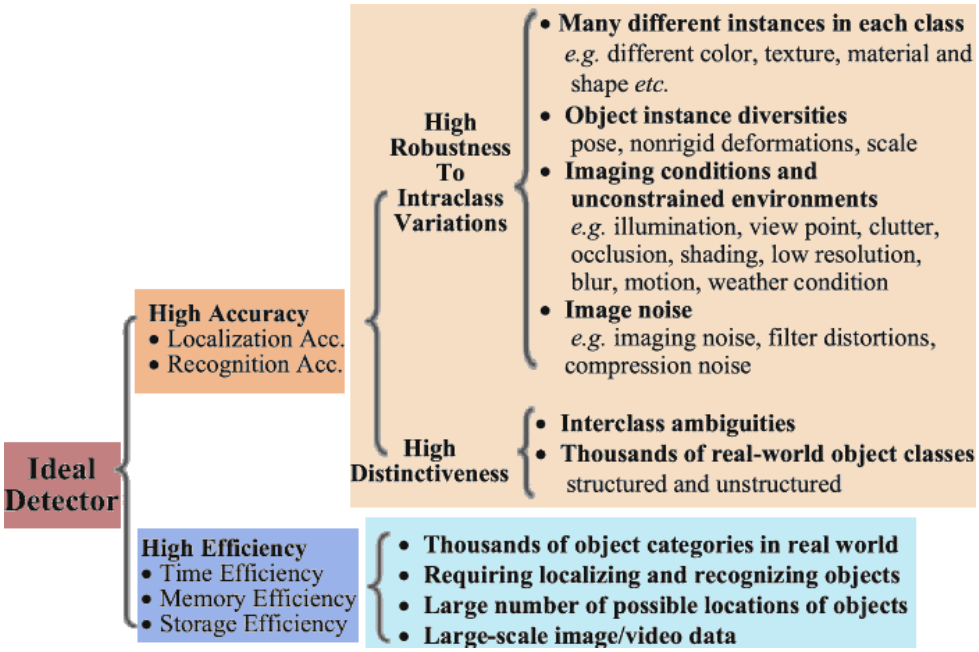

Figure 2: Classifications of the most common OD challenges.

The OD quality is determined by the precision, error, detection rates and time consumption of the OD systems in which a trade-off among these factors is essential. Generally, a complete OD system is comprised of the features, recognition, and localization. Features are of 2 types depending on the area (spatial) and method of acquisition. Furthermore, the features acquirement mode can be classified into 3 types including the extraction, selection, and composition [7]. These features may be local or global [29].

In general, the objects are recognized using the matching or learning methods. The matching technique recognizes the objects according to the distance between the target features and stored template. Conversely, the learning technique mimics the human ability to learn from the instances. It recognizes the unknown objects based on the earlier experienced and knowledge. The machine learning techniques are of 2 types of ensemble-based and single-based [11]. In the single-based methods, only one algorithm is applied to learn and classify the object instances [11]. However, in the ensemble-based methods numerous algorithms are used to obtain the final computational decisions for the instance's recognition [43, 44]. Figure 3 presents the core elements of a complete OD system and the sub-categories associated to each element.

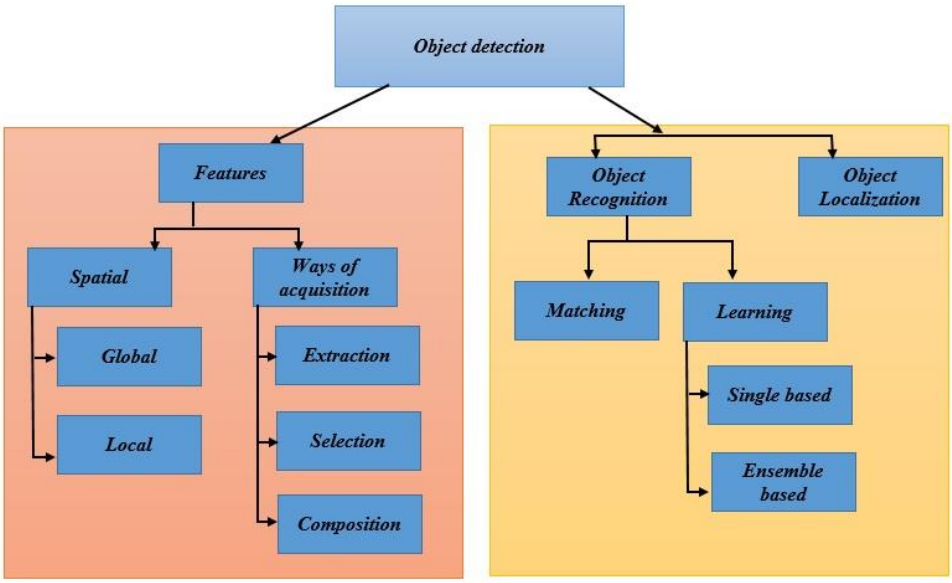

Figure 3: The main elements of a complete OD system.

Considering the immense fundamental and applied significance of the deep learning algorithms-based OD and localization in the CV, the present paper provides a comprehensive overview concerning the recent progress and future trends of this field. The performance of various state-of-the-art techniques, algorithms, models and methods used for the accurate detections and localizations of the objects are explained and compared. The salient features of these techniques and algorithms with their advantages and disadvantages in diverse applications are underscored. The working principle of the deep learning and the main features of the proposed algorithms are highlighted. The basic architectures of the standard OD systems are described. The detailed categorizations of the OD and localization approaches are presented.

2. Deep Learning Convolutional Neural Network

The local features of the data are generally extracted using a kind of artificial neural network (ANN) called the convolutional neural network (CNN). Through the allocations of the weights on the singular maps of the features it is possible to simplify the network models via the CNN, enabling a decrease in the total weights. These features made the CNN most celebrated model for the pattern recognitions. The CNN is used via the documents reading system that is trained together with a probability model consisted of the language restrictions.

Historically, this system was used in the US during late 1990s to read nearly 10% of the dispersed bank cheques. Afterward, several CNN-based optical handwriting and characters recognition systems were introduced by the Microsoft [45]. In the early 1990s, the CNNs were used as an experiment for the OD in the natural images such as the hands and faces. At the same period, CNNs were used to solve the issues related to the speech recognition [46] and document reading [47, 48] . Meanwhile, the time-delay neural networks (TDNNs) were used to extract the meaningful contents. The probabilistic model was combined with the CNN for the document reading of barriers that existed in the languages. Subsequently, it was extensively applied in the US to read the cheques. A deep learning (DL) algorithm was introduced by unifying the transfer learning and multiple tasks learning to analyze the images of the biological structures [49]. Later, [51] developed a CNN-based DL algorithm which outperformed the existing ML strategies. Since then, the CNN-based DL algorithm was widely exploited to resolve the visual recognition problems and became promising. In the viewpoint, it is important to discuss the basic architecture of the CNN.

2.1 CNN Architectures

The structural design of the CNN is constituted of three layers such as the input, hidden (latent), and output. The latent or hidden layers are called the pooling or fully-connected or convolutional layers. Figure 4 shows the basic architecture of the CNN [52]. The next section provides a brief description of this layer.

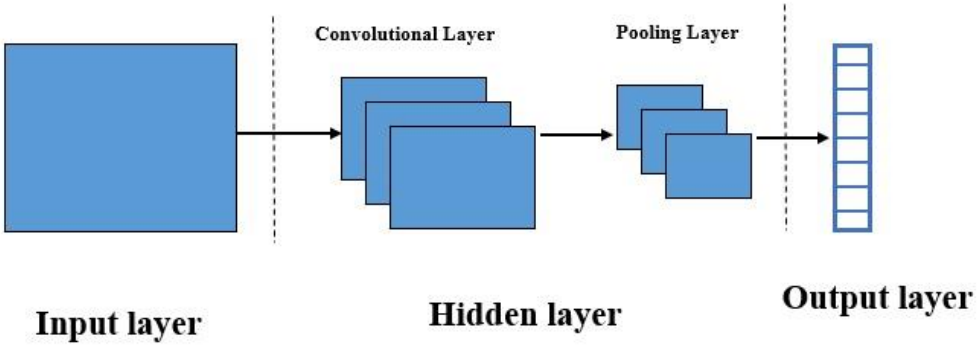

Figure 4: The schematics of the basic architectures of CNN [52]

I. Convolutional Layer

This is principal layer in the structure of the CNN. The given functions are executed iteratively via the convolution processes to generate a changing output function [53]. This convolutional layer is composed of many maps of the neurons called the maps of the filters or features. Regarding the size, it is somewhat similar to the dimension of the input data. The neural reactivity can be interpreted by quantifying the discrete convolution of the receptors. The quantification includes the calculations of the total neural weights of the input and the assignments of the activation function. Figure 5 depicts the structure of the typical discrete convolutional layer.

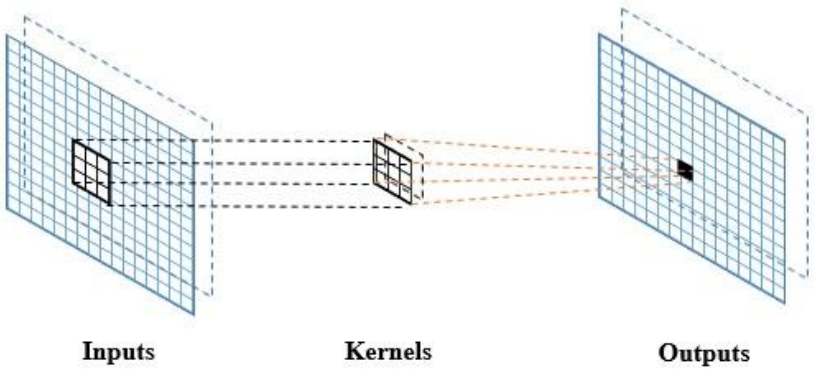

Figure 5: The architecture of the discrete convolutional layer [52].

II. Max Pooling Layer

The max pooling layer involves the production of many meshes obtained from the output of the segmented convolutional layer. The maximum grid value undergoes a sequencing in the matrices [52]. The operators are used for the calculation on each matrix so that the average or maximum value can be quantified. Figure 6 displays the construction of the max pooling layer.

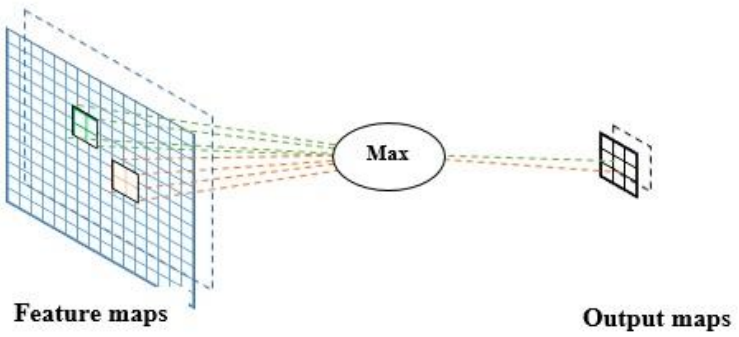

Figure 6: The construction of the max pooling layer [52].

III. Full Connection Layer

This layer refers to the practically whole CNN that containing 90% of the total CNN structural parameters. This layer allows the transmission of the input along the networks with the pre-configured vector length [51]. In this network, the given data undergoes the transformation through a layer before being classified. In addition, the convolutional layer is also transformed, allowing the retention of the information integrity. The neurons from every previous layer are used to achieve the full connection layers. These fully connected layers are used as the final network layer and are involved in the classification. Figure 7 shows the configuration of a full connection layer. Figure 8 presents a typical complete CNN containing all the 3 layers.

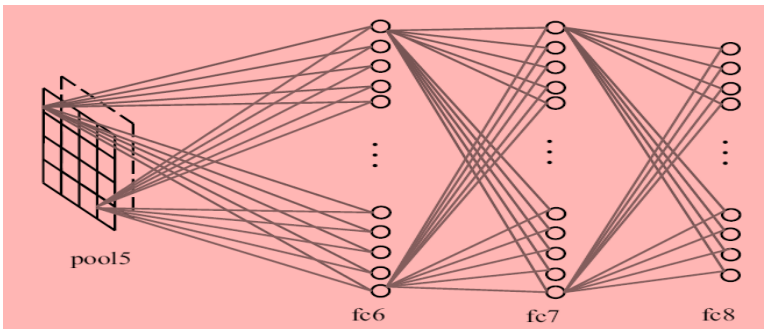

Figure 7: The configuration of the full connection layers [51].

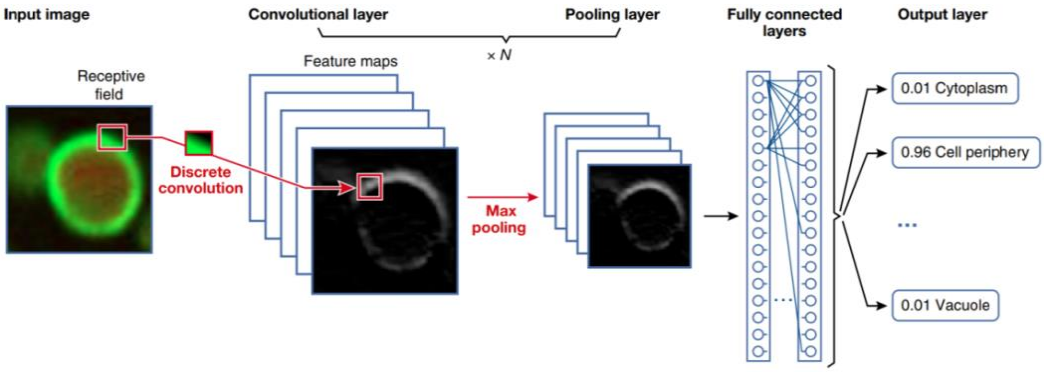

Figure 8: The typical architecture of a complete CNN.

It is important to note that the typical architecture of the complete CNN that demonstrated here may not be the best choice to solve the CV problems because it is designed for the object recognition. This it is necessary to design a customized network structure adaptive to the problem domain for better performance. However, the experimental results reveal that the constructed CNN can achieve the desired performances.

2.2 Region-based Convolutional Network

The region-based convolutional network (R-CNN) model begins by searching the region followed by the classifications. It uses the selective search (SS) technique [54] meticulously in an image to determine the position of the object where the small regions in the image are initialized and then merge via the ranked groups. Eventually, the group is in the form of a box that encloses the whole image. The identified regions are unified in accordance to the diversity of the color space and similarity matrix. Essentially, the output becomes a proposal with few regions that may enclose an object by unifying small regions as indicated in Figure 9.

Figure 9: Selective search applications of the algorithm with the visualization of the (top) segmentation and (bottom) region proposals [54].

The SS technique and DL are combined in the R-CNN model [55] for detecting the proposals of the regions and identifying the object in those detected regions, respectively. Every proposal of the regions is resized for matching the CNN input and extracting the features vector of dimension 4096 before being fed this vector into the multiple classifiers for producing the probabilities belonging to every class. Every such class possesses a SVM classifier that is trained to understand the probability for detecting the targeted object of a given features vector. Furthermore, such vector feeds the linear regression to familiarize the shapes of the bounding box of the region proposal, thereby reducing the localization error. The presented CNN model [55] was trained on the ImageNet dataset of 2012 for the image classifications. It was tuned finely via the proposals of the regions for the IoU above 0.5 with the boxes of ground-truths. Two forms were generated where one of them used the 2012 PASCAL VOC dataset and the other used the 2013 ImageNet dataset with the bounding boxes. In addition, the SVM classifiers were trained for every class of individual dataset. The best R-CNN model could achieve the mAP score of 62.4% (22.0 points enhancement respecting the second best result on the leader board) and 31.4% (7.1 points enhancement respecting the second best result on the leader board) for the former and later dataset, respectively. Figure 10 displays the architecture of a typical R-CNN.

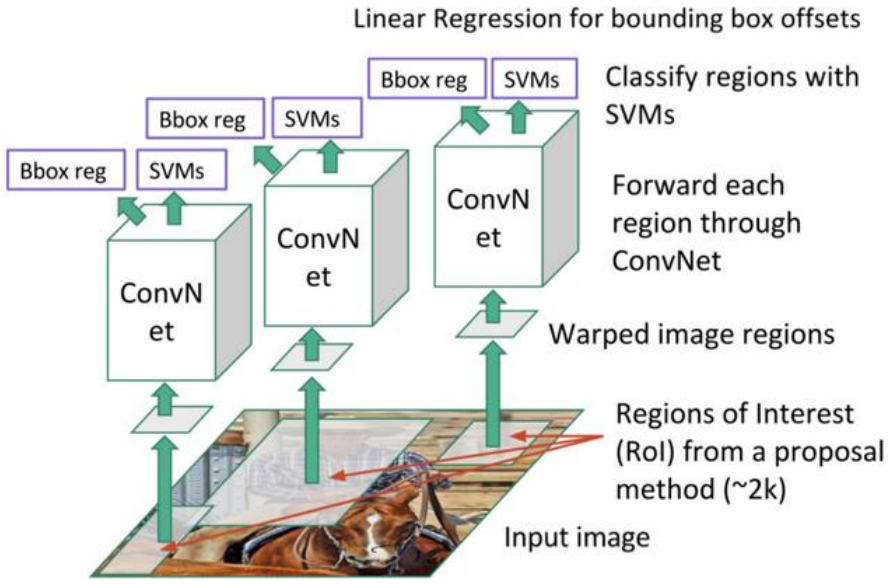

Figure 10: The construction design of R-CNN [56].

2.3 Fast Region-based Convolutional Network (FR-CNN)

The main reason of using the FR-CNN is to decrease the time taken for analyzing the proposals of all the regions using large number of models [57]. Rather than utilizing one CNN for every region proposals a leading CNN with many convolutional layers can be used to take the whole image as the input. The SS technique can be applied on the obtained regions of the feature maps for detecting the region of interests (ROIs). The size of the feature maps become less when a pooling layer is used to obtain the valid ROIs with constant hyper-dimension (height and width). Every ROI layer feeds full connection layers to create the vector of features. This vector is further utilized in predicting the perceived object with a softmax classifier, familiarizing the localizations the bounding box with the linear regression. Figure 11 shows the basic architecture of the FR-CNN. The best FR-CNN yields a mAp score of 70.0%, 68.8% and 68.4% for the 2007 PASCAL VOC, 2010 PASCAL VOC and 2012 PASCAL VOC test dataset, respectively.

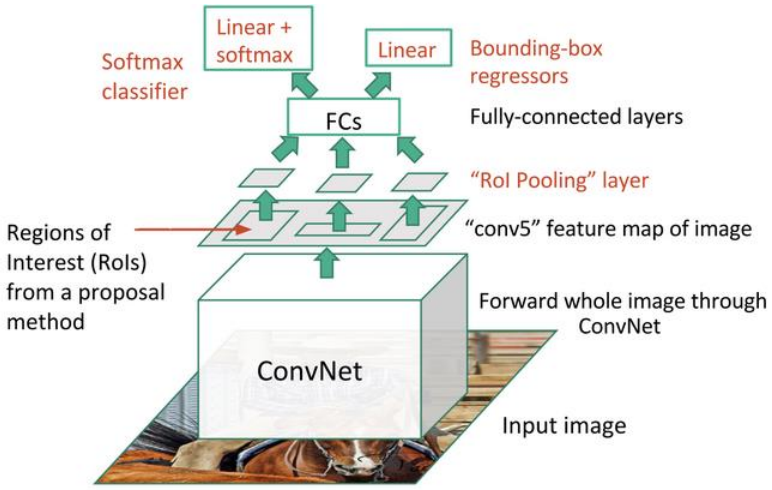

Figure 11: The basic framework and working principle of the FR-CNN [57].

2.4 Faster Region-based Convolutional Network (Faster-R-CNN)

The proposals of the region detected using the SS technique is further needed by the preceding model, thus increasing the computational expense. To overcome this problem, [58] proposed a region proposal network (RPN) for the direct creation of the region proposals, prediction of the bounding boxes and detection of the objects. The Faster-R-CNN is the unification between the RPN and FR-CNN models. The CNN model accepts the whole image as the input to produce the feature maps. A sliding window of size 3 × 3 for the whole feature maps and output of the features vector are interlinked to the two full connecting layers with one for the regressions box and another for the classifications box. The proposals for the multiple regions are predicted using the full connection layers where a maximum of k regions is fixed. Therefore, the size of the output of the regressions and classifications box layers is 4 k and 2k, respectively. Figure 12 explains detection scheme of the k region proposals detected using the sliding window known as the anchor boxes.

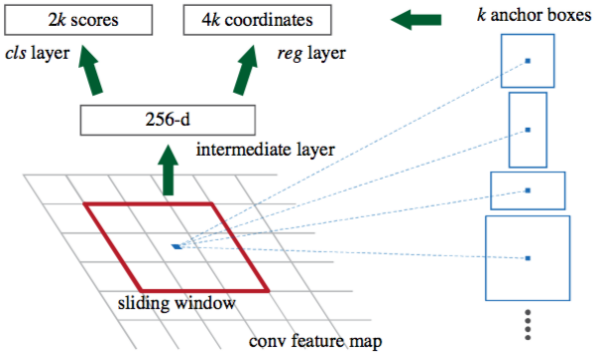

Figure 12: The detection of the anchor boxes for a single 3 × 3 windows [58].

The anchor boxes upon the detection are chosen via a threshold according to their objectness scores to retain the appropriate boxes only. The anchor boxes and feature maps obtained using the first CNN model is fed to a FR-CNN. The Faster-R-CNN utilizes the RPN to evade the SS technique, thereby improving the performance by accelerating the training and testing process. The RPN utilizes a pre-trained model on the ImageNet dataset for the classifications and the fine tuning is performed on the PASCAL VOC dataset. Finally, the created regions of the proposals with the anchor boxes are utilized the Fast-R-CNN training iteratively. The best Faster-R-CNNs produced a mAP score of 78.8% and 75.9% when tested on the 2007 PASCAL VOC and 2012 PASCAL VOC dataset after the training via

the PASCAL VOC and COCO datasets. One of the models was shown to be faster by a factor 34 compared to the FR-CNN. Figure 13 displays the basic configuration of the SS technique revealing the feeding of a CNN model by the image feeds to create the anchor boxes as the proposals of the region with the assurance of enclosing an object. In addition, FR-CNN is utilized to input the feature maps and proposals of the region where every box produced the likelihoods for detecting every object and correcting the box position.

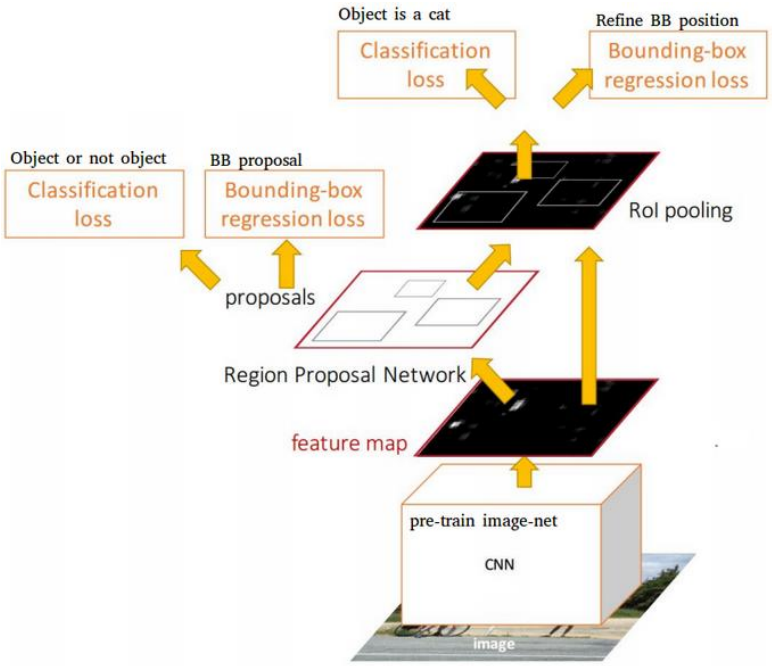

Figure 13: Basic configuration of the SS technique revealing the working [59].

2.5 *Region-based Fully-Convolutional Network (R-FCN)*

The FR-CNN and Faster-R-CNN techniques are used to detect the proposals of the region and identify an object in the individual region. The R-FCN [60] model has a single convolutional layer that enables the whole back-propagation for the training and inferring. In this model, two basic phases are combined to consider concurrently the OD (location invariants) and its position (location variants). The ResNet-101 model receives the first image as the input and the final layer is used to output the feature maps, where every layer can detect a class at some position. For instance, a particular feature map is specified to detect the cat, other one is specialized to detect the banana and so forth. These kinds of feature maps are known as the location-sensitive score maps due to their spatial localization traits of a specific object. These maps have dimension of $k \times k \times (C+1)$ where k and C are the size and classes number, respectively. Such maps constitute the scores bank wherein patches can be created to identify an object's portion. For instance, the 3×3 parts of an object can be recognized with k=3. In addition, a RPN must be run separately to generate the ROI and each of them has to be eventually cut into bins to check them against the scores bank. The patches vote turn out to be 'yes' for the activation of sufficient of these parts. Figure14 shows the creation of feature maps using the ResNet-101 model and object recognition protocols via the R-FCN model.

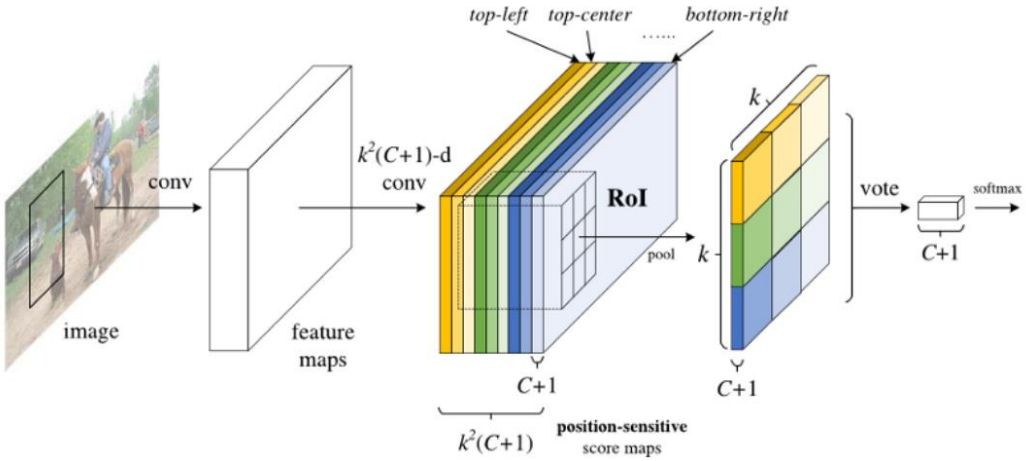

Figure 14: Creation of the feature maps, detection of the ROI via the RPN model and scores computation for every region to decide the presence of the most probable object [60].

The R-FCN for the person class with size 3 × 3 when the ROI is correctly and incorrectly overlapped the object, respectively [60]. Clearly, the R-FCN model could meticulously detect a person when the ROI is correctly overlapped at the image center. The sub-regions of the feature maps are precise to the patterns related to the person under visualization, thereby voting yes to signify the presence of a person at that position. Conversely, when the ROI is incorrectly overlapped at the image center or moved on the right hand side the visualization is remarkably affected. In case of non-overlapping region, the sub-regions of the feature maps are completely disagreed to detect the person, thereby voting vote no to indicate the complete absence of person at tht position. The best R-FCN produced the mAP score of 83.6% and 82.0% when tested on the 2007 PASCAL VOC and COCO datasets, respectively. In addition, it yield the score of 53.2% and 31.5% when tested on the test-dev 2015 COCO challenge (with IoU = 0.5) and official mAP metric datasets, respectively. It is asserted that the R-FCN model is faster by the factors of 2.5 to 20 compared to its Faster R-CNN model counterpart.

2.6 *You Only Look Once (YOLO) Model*

Using this model it is possible to predict directly in real time the bounding boxes and the likelihood of the classes for one network in a particular assessment [61]. First, the YOLO model accepts an image as its input and then the image is divided in the grids of dimension S × S where its every cell can predict the number of the bounding boxes (B) with a precise confidence score. This score is defined as the likelihood of detecting the object multiplied by the IoU amid the boxes that are anticipated and ground truth ones as illustrated in Figure 15.

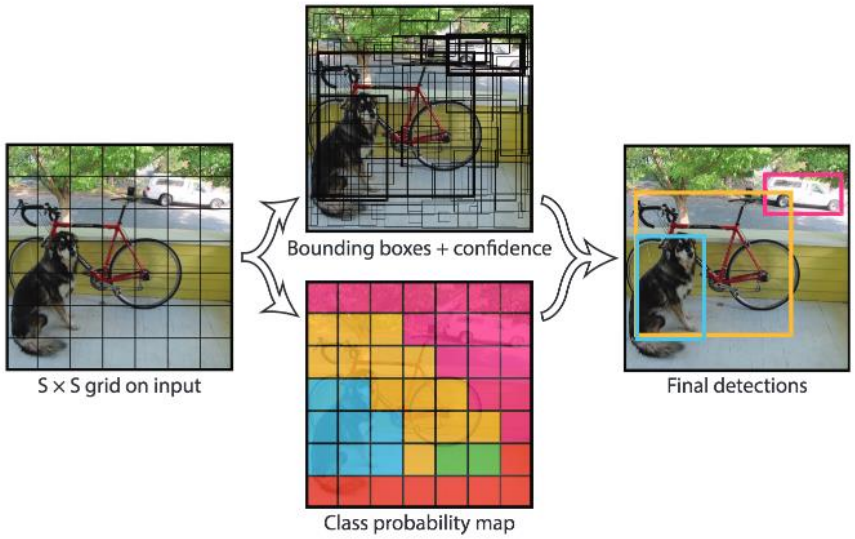

Figure 15: The working of the YOLO model showing the input image division in grids, predicted bounding boxes and classes with the confidence scores so called the regression and classification, respectively [61].

The main motivation of using the CNN model came from the GoogLeNet that is introduced at the foundation modules. This network possesses twenty four convolutional subsequently two full connection layers. The reducing layers are comprised of the filters of dimension 1 × 1 with subsequent convolutional layers of size 3 × 3 replacing the original inception modules. The fast YOLO model is the simpler one that contains nine convolutional layers and few filters where majority of the convolutional layers is pre-trained via the ImageNet dataset with the classification. Four convolutional layers with subsequent 2 full connection layers are combined with the earlier network and totally re-trained with the 2007 and 2012 PASCAL VOC datasets. The tensor of dimension S × S × (C + B × 5) is outputted by the last layer matching the prediction of every grid's cell (where B and C denotes the calculated number of anchor boxes per cell and chances). Herein, every box is linked to 4 coordinates such as the box center, width, height and confidence score. The bounding boxes predicted by the earlier models normally enclosed an object, but the YOLO model could predict the large number of the bounding boxes and several of the devoid of any object.

Figure 16 illustrates the structural configuration of the YOLO model that used the non-maximum suppression (NMS) method at the network, comprising of extremely overlapped bounding boxes of the same object merged in one despite the presence of few false positive. The YOLO model produced the mAP score of 63.7% and 57.9% (with real time performance) when tested on the 2007 and 2012 PASCAL VOC datasets.

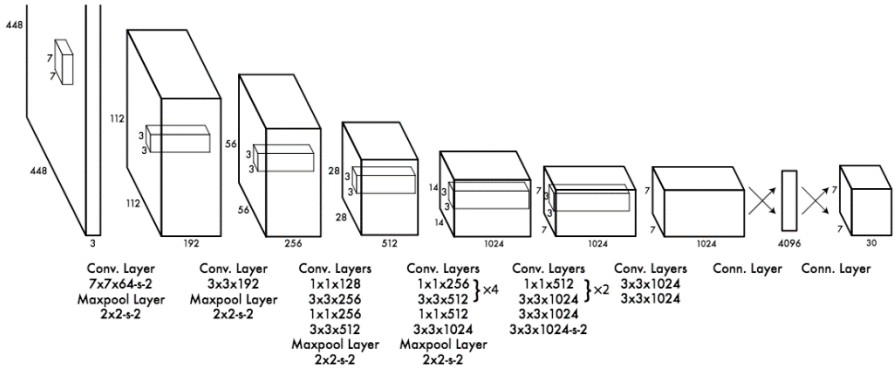

Figure 16: The structural configuration of the YOLO model with two full connections and twenty four convolutional layers, respectively [61].

Table 1 shows the comparative evaluations of the existing state-of-the-art techniques for the CV with their added advantages and disadvantages. The used datasets and the main results of each report with their limitations are emphasized.

Table 1: Comparative evaluations of the existing literatures revealing the merits and demerits of different methods used in the CV.

References	Dataset Type	Methods Used	Main Outcomes
Janssens et al (2017)[62]	Thermal Images	CNN model to detect the faulty parts and estimate the oils level in the machinery.	Produced the score of 95% (detection) and 91.67% (estimation).
Rodin et al. (2018)[63]	Thermal images	GMM and CNN model to detect and classify the objects.	Yield 92.5% accuracy over the testing dataset.
Suhao et al. (2018)[64]	RGB	Improved Faster-RCNN model for the objects detection.	Provided good results.
Nam et al. (2018)[65]	Thermal and RGB images	Features extraction with contrast, entropy, energy and homogeneity.	Precision for the visible spectral images of 92.7% and thermal images of 65.8%.
Masita et al. (2019)[66]	RGB images	R-CNN model to detect the pedestrians.	Produced 52% precision on the Penn-Fudan dataset and 84% on the KTH football dataset.
Akula et al. (2019)[67]	Thermal images	CNN model with various data augmentation techniques for the detection of potholes.	Pre-trained ResNet-152 CNN model yield a precision of 97.08%
Xu et al. (2019)[68]	Thermal images	Faster-RCNN and ACF-based methods to detect the pedestrians.	Recommended the methods such as sample balance handling, anchor setting, backbone selection and Faster R-CNN with higher features resolution to get improved performance.

3. Conclusions

This paper comprehensively overviewed the overall developments of CV fields with many potential applications related to the DL algorithms-based ODs and localizations. Lately, DL-based ODs have shown unbound promises owing to their robust learning capacity and benefits to deal the occlusions, scale transformations and background switching. With ever-increasing growth and demands the OD technologies are expected to localize and classify all simultaneously to accomplish a totally distinctive network. Therefore, it would be possible to train from the top to the bottom with the back-propagation. Many recent models have made the trade-off for high performing prediction capacity. The CV area is in its infancy and many new avenues are yet to open up with intensive studies to get a better insight of the OD landscapes. In short, this article provided the detailed progression of the neural networks and associated learning systems, serving as the taxonomy for navigating the CV field.

References

[1] Adams, A. A., & Ferryman, J. M. (2015). The future of video analytics for surveillance and its ethical implications. *Security Journal*, *28*(3), 272-289.

[2] Nixon, M., & Aguado, A. (2019). *Feature extraction and image processing for computer vision*. Academic Press.

[3] Schmiedel, J. M., Klemm, S. L., Zheng, Y., Sahay, A., Blüthgen, N., Marks, D. S., & van Oudenaarden, A. (2015). MicroRNA control of protein expression noise. *Science*, *348*(6230), 128-132.

[4] Pflüger, H., & Ertl, T. (2016). Sifting through visual arts collections. *Computers & Graphics*, *57*, 127-138.

[5] DiCarlo, J. J., & Cox, D. D. (2007). Untangling invariant object recognition. *Trends in cognitive sciences*, *11*(8), 333-341.

[6] Felzenszwalb, P. F., Girshick, R. B., & McAllester, D. (2010, June). Cascade object detection with deformable part models. In *2010 IEEE Computer society conference on computer vision and pattern recognition* (pp. 2241-2248). IEEE.

[7] Savitzky, A. H., Mori, A., Hutchinson, D. A., Saporito, R. A., Burghardt, G. M., Lillywhite, H. B., & Meinwald, J. (2012). Sequestered defensive toxins in tetrapod vertebrates: principles, patterns, and prospects for future studies. *Chemoecology*, *22*(3), 141-158.

[8] Gupta, A. K., Sharma, M., Khosla, D., & Singh, V. (2019). Object detection of colored images using improved point feature matching algorithm. CENTRAL ASIAN JOURNAL OF MATHEMATICAL THEORY AND COMPUTER SCIENCES, 1(1), 13-16.

[9] Barczak, A. L. C. (2007). *Feature-based rapid object detection: from feature extraction to parallelisation: a thesis presented in partial fulfilment of the requirements for the degree of Doctor of Philosophy in Computer Sciences at Massey University, Auckland, New Zealand* (Doctoral dissertation, Massey University).

[10] Soofi, A. A., & Awan, A. (2017). Classification techniques in machine learning: applications and issues. *Journal of Basic and Applied Sciences*, *13*, 459-465.

[11] Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, *160*, 3-24.

[12] Taylor, M. E., & Stone, P. (2009). Transfer learning for reinforcement learning domains: A survey. *Journal of Machine Learning Research*, *10*(Jul), 1633-1685.

[13] Ansari, M. A., & Singh, D. K. (2018, March). Review of Deep Learning Techniques for Object Detection and Classification. In *International Conference on Communication, Networks and Computing* (pp. 422-431). Springer, Singapore.

[14] Galleguillos, C., Rabinovich, A., & Belongie, S. (2008, June). Object categorization using co-occurrence, location and appearance. In *2008 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1-8). IEEE.
[15] Lampert, C. H., Blaschko, M. B., & Hofmann, T. (2008, June). Beyond sliding windows: Object localization by efficient subwindow search. In *2008 IEEE conference on computer vision and pattern recognition* (pp. 1-8). IEEE.
[16] Tien, N. T., Jeon, S., Kim, D. I., Trung, T. Q., Jang, M., Hwang, B. U., ... & Bao, Z. (2014). A flexible bimodal sensor array for simultaneous sensing of pressure and temperature. *Advanced Materials*, *26*(5), 796-804.
[17] Heitz, G., Elidan, G., Packer, B., & Koller, D. (2009). Shape-based object localization for descriptive classification. In *Advances in Neural Information Processing Systems* (pp. 633-640).
[18] Harzallah, H., Jurie, F., & Schmid, C. (2009, September). Combining efficient object localization and image classification. In *2009 IEEE 12th international conference on computer vision* (pp. 237-244). IEEE.
[19] Zhang, Z., Cao, Y., Salvi, D., Oliver, K., Waggoner, J., & Wang, S. (2010, June). Free-shape subwindow search for object localization. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (pp. 1086-1093). IEEE.
[20] Dai, Q., & Hoiem, D. (2012, June). Learning to localize detected objects. In *2012 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3322-3329). IEEE.
[21] Zhu, P., Wen, L., Du, D., Bian, X., Ling, H., Hu, Q., ... & Liu, X. (2018). VisDrone-VDT2018: The vision meets drone video detection and tracking challenge results. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 0-0).
[22] Du, D., Qi, Y., Yu, H., Yang, Y., Duan, K., Li, G., ... & Tian, Q. (2018). The unmanned aerial vehicle benchmark: Object detection and tracking. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 370-386).
[23] Ahsan, A. M., & Mohamad, D. B. (2017). Machine learning technique for object detection based on SURF feature. International Journal of Computational Vision and Robotics, 7(1-2), 6-19.
[24] Barczak, A. L. (2007). Feature Based Rapid Object Detection: From Feature Extraction to Parallelisation. Doctor of Philosophy. Massey University, Auckland, New Zealand.
[25] Alwaili, S. (2007). Image-based object detection and identification. Doctor Philosophy, Universiti Teknologi Malaysia, Faculty Geoinformation Science and Engineering.
[26] Ni, Z.-S. (2012). B-SIFT: A Binary SIFT Based Local Image Feature Descriptor. Fourth International Conference on Digital Home (ICDH), Guangzhou, China,117-121.
[27] Schmiedel, T., Einhorn, E., & Gross, H. M. (2015, September). IRON: A fast interest point descriptor for robust NDT-map matching and its application to robot localization. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 3144-3151). IEEE.
[28] Calonder, M., Lepetit, V., Ozuysal, M., Trzcinski, T., Strecha, C., and Fua, P.(2012c). BRIEF: Computing a local binary descriptor very fast. IEEE Transactions on Pattern Analysis and Machine Intelligence. 34(7), 1281-1298.
[29] Pflüger, H., & Ertl, T. (2016). Sifting through visual arts collections. Computers & Graphics, 57, 127-138.
[30] Nixon, M., & Aguado, A. (2019). Feature extraction and image processing for computer vision. Academic Press.
[31] Miller, I., Campbell, M., Huttenlocher, D., Kline, F. R., Nathan, A., Lupashin, S., et al. (2008). Team Cornell's Skynet: Robust perception and planning in an urban environment. Journal of Field Robotics. 25(8), 493-527.
[32] Kim, D., and Dahyot, R. (2008). Face components detection using SURF descriptors and SVMs.

International Conference on Machine Vision and Image IMVIP, Portrush, 51-56.
[33] Szeliski, R. (2010). Computer vision: algorithms and applications: Springer. Tamminen, T., and Lampinen, J. (2004). A Bayesian occlusion model for sequential object matching. Proceeding of the British Conference on Machine Vision Conference, 547-556.
[34] Felzenszwalb, P. F., Girshick, R. B., McAllester, D., and Ramanan, D. (2010). Object Detection with Discriminatively Trained Part-Based Models. IEEE Transactions on Pattern Analysis and Machine Intelligence, 32(9), 1627-1645.
[35] Mikolajczyk, K., and Schmid, C. (2004a). Scale and affine invariant interest point detectors. International journal of computer vision. 60(1), 63-86.
[36] Favelle, S. K., Palmisano, S., and Avery, G. (2011). Face viewpoint effects about three axes: The role of configural and featural processing. Perception- London. 40(7), 761.
[37] Bay, H., Ess, A., Tuytelaars, T., and Van Gool, L. (2008). Speeded-up robust features (SURF). Computer Vision and Image Understanding. 110(3), 346-359.
[38] Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. International journal of computer vision. 60(2), 91-110.
[39] Surasak, T., Takahiro, I., Cheng, C. H., Wang, C. E., & Sheng, P. Y. (2018, May). Histogram of oriented gradients for human detection in video. In 2018 5th International Conference on Business and Industrial Research (ICBIR) (pp. 172-176). IEEE.
[40] Kou, Q., Cheng, D., Chen, L., & Zhao, K. (2018). A multiresolution gray-scale and rotation invariant descriptor for texture classification. IEEE Access, 6, 30691-30701.
[41] Chaudhury, A., & Barron, J. L. (2018, May). Occluded leaf matching with full leaf databases using explicit occlusion modelling. In 2018 15th Conference on Computer and Robot Vision (CRV) (pp. 8-15). IEEE.
[42] O'Mahony, N., Campbell, S., Carvalho, A., Krpalkova, L., Hernandez, G. V., Harapanahalli, S., ... & Walsh, J. (2019). One-Shot Learning for Custom Identification Tasks; A Review. Procedia Manufacturing, 38, 186-193.
[43] Shadman Roodposhti, M., Lucieer, A., Anees, A., & Bryan, B. A. (2019). A Robust Rule-Based Ensemble Framework Using Mean-Shift Segmentation for Hyperspectral Image Classification. Remote Sensing, 11(17), 2057.
[44] Galar, M., Fernández, A., Barrenechea, E., Bustince, H., and Herrera, F. (2012). A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, 42(4), 463-484.
[45] Park, E. *et al.* (2016) 'Combining Multiple Sources of Knowledge in Deep CNNs for Action Recognition', *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 53, pp. 1–8. doi: 10.1109/WACV.2016.7477589.
[46] Atzori, M., Cognolato, M. and Müller, H. (2016) 'Deep learning with convolutional neural networks applied to electromyography data: A resource for the classification of movements for prosthetic hands', *Frontiers in Neurorobotics*, 10(SEP), pp. 1–10. doi: 10.3389/fnbot.2016.00009.
[47] Gholamrezazadeh, S., Salehi, M. A. and Gholamzadeh, B. (2009) 'A comprehensive survey on text summarization systems', *Proceedings of the 2009 2nd International Conference on Computer Science and Its Applications, CSA 2009*. doi: 10.1109/CSA.2009.5404226.
[48] Saggion, H. and Poibeau, T. (2016) 'Automatic Text Summarization: Past, Present and Future', *Multi-source, Multilingual Information Extraction and Summarization, Springer, pp.3-13,* pp. 3–13. Available at: https://hal.archives-ouvertes.fr/hal-00782442.
[49] Zeng, T. and Ji, S. (2016) 'Deep convolutional neural networks for multi-instance multi-task learning', *Proceedings - IEEE International Conference on Data Mining, ICDM*, 2016-Janua, pp. 579–588. doi: 10.1109/ICDM.2015.92.
[50] Liao, Q. *et al.* (2018) 'Cancer classification with multi-task deep learning', *2017 International*

Conference on Security, Pattern Analysis, and Cybernetics, SPAC 2017, 2018-Janua, pp. 76–81. doi: 10.1109/SPAC.2017.8304254.

[51] Krizhevsky, A., Sutskever, I. and Hinton, G. E. (2012) 'ImageNet Classification with Deep Convolutional Neural Networks', *Advances In Neural Information Processing Systems*, pp. 1–9. doi: http://dx.doi.org/10.1016/j.protcy.2014.09.007.

[52] Liu, Y. and An, X. (2018) 'A classification model for the prostate cancer based on deep learning', *Proceedings - 2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics, CISP-BMEI 2017*, 2018-Janua, pp. 1–6. doi: 10.1109/CISP-BMEI.2017.8302240.

[53] O'Connor, P. *et al.* (2013) 'Real-time classification and sensor fusion with a spiking deep belief network', *Frontiers in Neuroscience*, 7(7 OCT), pp. 1–13. doi: 10.3389/fnins.2013.00178.

[54] Uijlings, J. R., Smeulders, A. W., & Scha, R. J. (2012). The visual extent of an object. *International journal of computer vision*, *96*(1), 46-63.

[55] Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 580-587).

[56] Kaul, S. (2018). Region Based Convolutional Neural Networks for object detection and recognition in ADAS application.

[57] Girshick, R. (2015). Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision* (pp. 1440-1448).

[58] Ren, S., He, K., Girshick, R., Zhang, X., & Sun, J. (2016). Object detection networks on convolutional feature maps. *IEEE transactions on pattern analysis and machine intelligence*, *39*(7), 1476-1481.

[59] Kang, M., Leng, X., Lin, Z., & Ji, K. (2017, May). A modified faster R-CNN based on CFAR algorithm for SAR ship detection. In *2017 International Workshop on Remote Sensing with Intelligent Processing (RSIP)* (pp. 1-4). IEEE.

[60] Dai, J., Li, Y., He, K., & Sun, J. (2016). R-fcn: Object detection via region-based fully convolutional networks. In *Advances in neural information processing systems* (pp. 379-387).

[61] Redmon, J., & Farhadi, A. (2016). YOLO9000: better, faster, stronger (2016). *arXiv preprint arXiv:1612.08242*, *394*.

[62] O. Janssens, R. V. d. Walle, M. Loccufier and S. V. Hoecke,"Deep Learning for Infrared Thermal Image Based Machine Health Monitoring," IEEE/ASME Transactions on Mechatronics, vol. Volume: 23, no. Issue: 1, pp. 151-159, 2017.

[63] Rodin, C. D., Lima, L. N., Andrade, F. A., Haddad, D. B., Johansen, T. A., & Storvold, R. (2018). Object Classification in Thermal Images using Convolutional Neural Networks for Search and Rescue Missions with Unmanned Aerial Systems.International Joint Conference on Neural Networks (IJCNN) , 1-8.

[64] Suhao, L., Jinzhao, L., Guoquan, L., Tong, B., Huiqian, W., & Yu, P. (2018). Vehicle type detection based on deep learning in traffic scene. *Procedia computer science*, *131*, 564-572.

[65] Nam, Y., & Nam, Y. C. (2018). Vehicle classification based on images from visible light and thermal cameras. *EURASIP Journal on Image and Video Processing*, *2018*(1), 1-9.

[66] Masita, K. L., Hasan, A. N., & Paul, S. (2018, November). Pedestrian detection using R-CNN object detector. In *2018 IEEE Latin American Conference on Computational Intelligence (LA-CCI)* (pp. 1-6). IEEE.

[67] Akula, A., & Sardana, H. K. (2019, October). Deep CNN-based Feature Extractor for Target Recognition in Thermal Images. In *TENCON 2019-2019 IEEE Region 10 Conference (TENCON)* (pp. 2370-2375). IEEE.

[68] Xu, Z., Zhuang, J., Liu, Q., Zhou, J., & Peng, S. (2019). Benchmarking a large-scale FIR dataset for on-road pedestrian detection. *Infrared Physics & Technology*, *96*, 199-208.