Stationary Object Detection using RetinaNet and Kalman Filter

Kunapaneni Sai Ajay Kumar*, Yanamala Maneesha Reddy*, Kilaru Babji*, Chitturi Sai Naveen Kumar*,
Man Pawan Aditya*, Dinesh Naraharasetty*, Usha Kumari* and Shuvendu Rana†
Department of Computer Science and Engineering
SRM Universiy AP, Andhrapradesh India 522240
Email: *{kunapaneni_sai,yanamala_maneesha,kilaru_babji,chitturi_sai,man_pawan,
dinesh_naraharasetty,usha_kumari}@srmap.edu.in
†shuvendu@ieee.org

2022 International Conference on Intelligent Controller and Computing for Smart Power (ICICCSP) | 978-1-6654-7258-6/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICICCSP53532.2022.9862392

***Abstract*—Detection of objects is the most popular research topic nowadays. In this regard Convolution neural network gives a direction to achieve the goal. But detection of the Stationary objects on a live camera become more challenging due to the non-rigid movement of the object. Also, most of the time stationary objects appear to be focal loss in the time of detection. So using CNN for those cases will make the scheme fragile. In this paper, Image segmentation and Kalman filter are used to rectify the focal loss to make the scheme more accurate. Here RetinaNet is used for the implementation of a better object detection scheme. As a result, it is observed that the use of RetinaNet makes the stationary object detection more accurate and the results are acceptable compared to the state of the art model.**

***Index Terms*—Object detection, Kalman Filter, RetinaNet, image, Video, Machine Learning**

I. Introduction

Object discovery has been a prominent exploration heading as well as a focus in PC vision, which can be used in programmed vehicles, mechanical technology, video observation, and pedestrian identification [1]. Openness of profound learning innovation has changed the customary methods of item distinguishing proof and article identification. The profound brain network has a lively element portrayal limit in picture handling and is generally utilized as the component extraction module in object location. Deep learning models don't require any special hand-drawn features and may be used as a classifier and relapse device [2] [3] [4] [5] [6]. As a result, deep learning innovation is an important prospect in object recognition. The issue articulation of article recognition is to figure out where items are really situated in a given picture (object confinement) and distinguish it. So the pipeline of customary article discovery models is primarily separated into three phases: useful area choice, highlight extraction, and recognition.

So we are going to use deep learning concepts. It has the advantages of a faster computation time, more accuracy, and ease of use. Object detection has made extensive use of deep learning technologies. RetinaNet [7] [8] [9] [10] is the most accurate deep learning concept. So this hybrid deep retina net has new approach to future object detection in which feature selection, extraction and classifier training can be performed

- Informative region selection: As different articles appear in various locations throughout the image and have varying viewpoint proportions or sizes. Filtering the entire image with a multi-scale sliding window is a must-do task. Albeit this thorough methodology can figure out all potential places of the articles, its inadequacies are additionally self-evident. Because of a colossal number of applicant windows, it is computationally costly and delivers such a large number of repetitive windows. In any case, if by some stroke of good luck a chance of inadmissible districts in the given picture.

- Highlight Extraction: We wish to remove visual aspects that can give a semantic and forceful portrayal in order to recognise diverse products. This is due to the factors that can produce depictions of complicated cells in the human mind. Regardless, due to the wide range of looks, lighting conditions, and foundations,, it's challenging to physically plan a vigorous component descriptor for impeccably portraying a wide range of items.

- Discovery: To make the introductions more progressive and instructional for visual recognition, an identifier recognises an objective article from a wide variety of classifications.

One of the basic undertakings in Computer Vision is the Detection and Tracking of articles. Yet at the same time, presently, the issues connected with this are creating. For the programmed location of moving articles, a portion of the checking frameworks can't ready to observe the distinction, when the distinction in brilliance between the foundation and the it is little to move objects. The expenses are exceptionally high in these frameworks. The greater part of the past techniques just focused on recognizing the unpleasant areas of targets. The precise moving objective location can't be accomplished. It causes the outcome to be displayed with clamor and mends and the calculation time is likewise expanded. Numerous frameworks can't tackle basic arrangements like

All the authors have given equal contribution.
978-1-6654-7258-6/22/$31.00 ©2022 IEEE

Partial Occlusions and Cross Targets. Regardless of whether the incomplete impediments occurred, due to different items or the intersection of other moving articles, the frameworks can't can continue to follow their objective. The item following can be intricate, because of camera movement, loss of data brought about by the projection of the 3D world on a 2D picture, varieties of target scale, incomplete impediments, continuous handling necessities, mess, etc. Along these lines it is hard to make an effective tracker and furthermore we can't follow the article precisely.

II. Literature Survey

[11] The paper's major goal was to find the target object on the road for driverless cars. The authors primarily described a deep neural network-based approach for object detection. The algorithm took the technique of providing an image input and detecting the object within it. The back end, front end, and side part information is retrieved from the input photos, and the object detection is performed utilising these extracted information as input to the deep learning algorithm. The algorithm's workflow starts with data collection. Machine learning concepts necessitate a training data set, so the author used the CIFAR-10 data set for training and imported it into the system to aid with feature extraction and target object detection.Then ground information like region of interest is provided to algorithm, so that it helps to detect the object by extracting the required features. Then depending upon the training data set and prediction of the algorithm the object is classified. The CNN was tweaked using 50,000 photos from the pre-trained data set. As a result, the author provided 40 on-road photos for algorithm verification, all of which were 100 percent accurate. The author has also developed a reformative CNN algorithm that lowers the costs associated with CNN algorithms. Furthermore, by raising the system's recognition rate, the system's accuracy was substantially improved.

[12] Autonomous vehicles require estimation of distance between moving objects. It requires image processing algorithms which can detect and classify objects without restricting the hardware platforms of the autonomous vehicle. Thus, author proposes use of Darknet based deep learning algorithm with modified detection stages. Darknet has various benefits like it is the mostly efficient real time detector. It has fast processing speed with simple network architecture. Kitty database is used to predict 3D bounding box from 2D convolution. To find the distance the bounding box values obtained from the image is used then manipulated with camera intrinsic and extrinsic parameters and finial transformation between coordinate systems give the absolute value

[13] Main functions of autonomous vehicle are to provide safety, which means the vehicle has to react suddenly when something appears in front. There are times where not only vehicles are present in front but pedestrians too it requires at most detection features to detect pedestrians accurately and reduced false detection. Thus detection of pedestrian can be considered as joint estimation between detection and semantic feature extraction by which the performance of the system can be increased. In the proposed method semantic it aims in assigning object class for each pixel of image. Signet architecture of deep learning was used for pixel wise classification. The encoder requires up-sampling stage for feature classification at lower resolutions. While performing up sampling to make use of pooling indices with corresponding encoders a decoder network was proposed. The architecture was trained using café -segment which has large dataset of urban traffic images. Where various semantic features like tree, bus, bi-cycle, pathways, few animals are labelled pixel wise and rest of the pixel remains blacked out at training stage. In next stage results of semantic features are integrated with detectors which make use of filtered channel features. Before feeding the input to decision forest stage for feature selection the output from these feature extraction channels is combined using sum-pooling method. Higher rate of fast and accurate classification can be achieved by using accumulated hard negative samples in cascade. This is done in booted forest method. Thus, integrated feature channels can be directly used to train booted forest as there is no constrain on dimension features. Thus, the results of the proposed method resulted in 4% more improvement by using more positive samples. More powerful detection of pedestrians depends upon the learning mechanism which asses the semantic cues.

A. FACTORS AFFECTING OBJECT DETECTOR

Object identification requires distinguishing the highlights that sway the exhibition of the finder with the system. In light of the writing overview the different variables which influence locator execution are include extractor, limit choice for misfortune computation, limit box encoding, preparing dataset, information expansion, restriction factors, and component planning layers. Object location utilizing CNN Deep Convolutional Neural Networks had been essentially utilized for object discovery. CNN is a sort of feed-forward brain organization and chips away at the statute of weight sharing. Convolution is a reconciliation showing how one trademark covers with another person and is a blend of these two capacities being increased. Fig. 4 shows the layered design of CNN for thing recognition. Picture is convolved with the actuation capacity to get the element extraction charts. To decrease the spatial intricacy of the organization, extraction maps are taken care of with inward layers to get disconnected highlight maps. This way is rehashed for the leaned toward no. of time and thusly, the component extraction maps are made. In the long run, those capacity maps are handled with totally related layers to get the result of picture acknowledgment showing a confidence rating for the expected class marks. For enhancing the intricacy of the organization and diminishing the wide assortment of boundaries, CNN utilizes particular types of pooling layers as displayed in table 1. Inward layers are interpretation invariant. Initiation maps are taken care of as a trigger to the pooling layers. They perform on each fix inside the chose map.

III. Design and Architecture

A. Overview

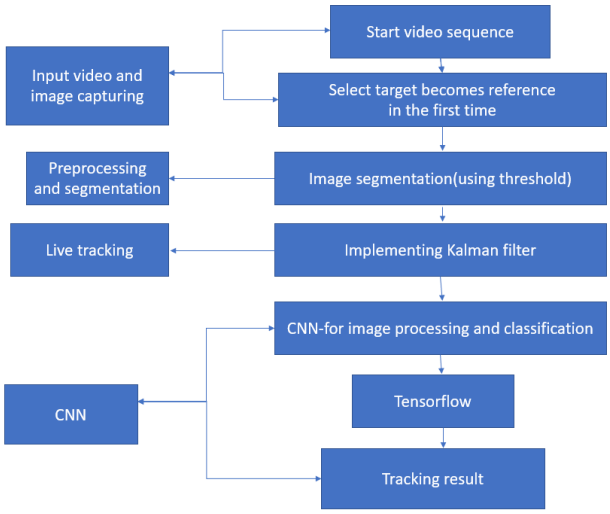


Fig. 1. Methodology Flow chart

The deep learning neural network has the advantages of being faster, more accurate, and easier to use. Object detection [14] has made extensive use of deep learning technologies. As a result, the temporal relationship of images is used in this paper to improve object detection accuracy. We used Kalman filter frameworks, we changed to improve the ability of tracking moving objects. The hybrid deep RetinaNet classifier is a new approach for feature reduction that combines feature selection, extraction, and training. At the same time, we used a genetic algorithm to combine this technique with RetinaNet and compare it to traditional feature, selection, and extraction techniques.

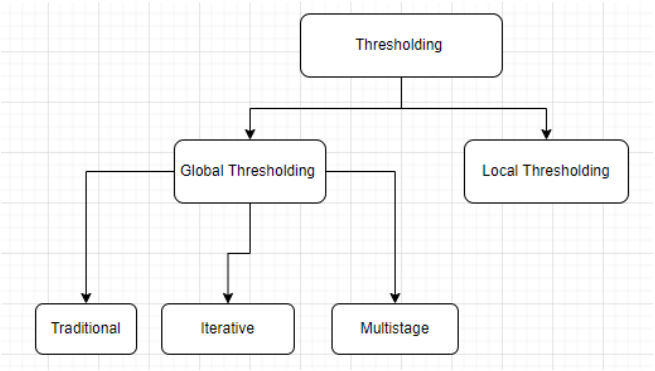
The object is placed in-front of the camera, Image segmentation is done by Thresholding and moving objects help that triggers kalman filter and is implemented simultaneously and then the picture is captured [15] [16] [17]. The classification and image processing is done by CNN. Object recognition using Tensorflow is a computer vision technology, this helps us recognize, identify, and track objects from images and videos. The final result is updated as the identified object from the given input data set.

B. Image Segmentation (Using Thresholding Technique)

One of the most significant approaches in image segmentation is the threshold technique.There are five steps in this technique:

- Create a grayscale version of the image.
- Adjust the image's threshold.
- Locate the contours of the image (edges).
- Using the biggest contour, create a mask.
- To delete the shadow, use the mask on the original image.

Thresholding



Thresholding [5] is used to convert a gray scale image to white pixels. Local adaptive thresholding differs from global thresholding. depending on its neighbours, selects separate threshold values for each pixel in the image. This technique can be expressed as:

$$T = T[x, y, p(x, y), f(x, y)] \tag{1}$$

The threshold value is T. The coordinates of the threshold value point are x, y. p(x,y) and f(x,y) are the grey level image pixels' points. It is possible to describe the threshold image g(x,y).

C. Kalman filter

A kalman filter is an algorithm for estimating the state of the system like live tracking of speed and position of objects using past and possibly noisy observations and current and possibly noisy measurements of that system. When the variables of interest can't be measured directly but an indirect measurement is available, Kalman filters are used to optimally estimate them as best as possible. When the measured values contain random error, uncertainty, unpredictability, or change, this technique uses a few equations and a few continuous data inputs to estimate the position, velocity, and other features of the object being measured.

Kalman filtering has so many kinds of technological applications in which most common application is that it is used for navigation, guidance, and control of vehicles. Kalman filtering is mostly applied in analysis of time series which is used for signal processing and econometrics. It is also mainly used for robotic motion planning, control and trajectory optimization. It will also be helpful for modeling of the CNS control movement.As a result of the time difference between the receiving motor commands from the brain and the receiving sensory feedback, Kalman filter's usage provides a realistic model for making approximate calculation of the current state of a system.

Kalman filter usually works by a two phase process. In the prediction phase, the Kalman filter finds the approximate calculations of the current state variables, besides their uncertainties. When the results of the following measurement

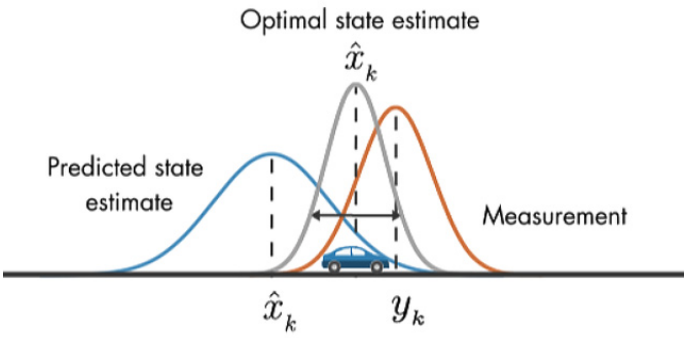


Fig. 2. Kalman filter

is noticed, these estimates are upgraded using a weighted average, with more weight given to these estimates with greater certainty.

Implementation in our project

- In this project, we create measurement matrix,
- Transition matrix and
- Process noise covariance matrix
- where all these kalman measurements of objects will be done on the camera frame appeared.
- All identified oberservations will be shown in blue colour and the kalman filter predictions will be shown in green colour.

D. Feature Extraction

Feature extraction is a process of converting the raw data into numerical features that can be used for particular problem while preserving the raw data in the original dataset. It helps providing better results than simply using machine learning technique to raw data.

Feature extraction can be done either by manually or automatically:

- Manual feature extraction requires you to identify and characterize features that are important to a particular problem and design a way to extract those features. Knowing the background and domain often helps you make informed decisions about which features are worthwhile. Engineers and scientists have created ways to extract information from photos, signals, and text after decades of research. The significance of the signal window is an example of a simple feature.
- Automatic feature extraction makes use of unique algorithms and deep networks to extract functions from indicators or photos with out the want for human intervention. This method may be very beneficial whilst you want to fast pass from uncooked information to constructing device getting to know algorithms. For instance, Wavelet scattering is an example of automatic feature extraction.

Feature extraction has basically been replaced by the first layer of deep networks with the advent of deep learning, but mainly for image data. Feature extraction is the first challenge for signal and time series applications and requires wide range of knowledge before building a good predictive model.

E. CNN

The convolutional neural network (CNN) is a type of deep, feed-forward artificial neural network that has been utilised to conduct accurate image categorization and identification in computer vision applications. Traditional neural networks are similar to CNNs, but with more layers. Through a nonlinear activation, it contains weights, biases, and outputs. The CNN's neurons are arranged in a volumetric manner, with height, width, and depth. The higher performance with image, speech, or audio signal inputs distinguishes convolutional neural networks from conventional neural networks. There are three primary types of layers, which are:

1) Convolutional layer
2) Pooling layer
3) Fully-connected (FC) layer

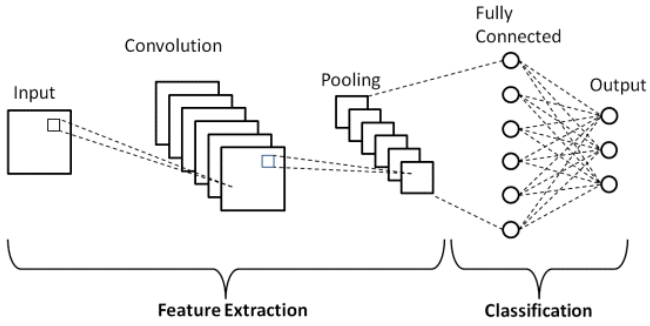


Fig. 3. CNN Architecture

F. Tensorflow object detection Model

When an image or video is displayed, TensorFlow object recognition works by using computer vision to identify known objects from all instances. The history of object recognition is as new as the Internet. The first neural network recorded for object detection was overwhelmed because the developers believed that object detection would help improve image recognition.

1) Why Choose TensorFlow Object Detection: Training a complete CNN from scratch can be time consuming and require a very large image dataset. There is a solution to this problem. The solution is to take advantage of TensorFlow's migration learning and object discovery APIs. You can use it to train, build, and deploy object discovery models. Fortunately, the API has several pre-trained models.

G. Retina Net

RetinaNet is made up of two enhancements to the single-stage object detection approach (like YOLO and SSD):

1) Feature Pyramid Networks for Object Detection
2) Focal Loss for Dense Object Detection

Feature Pyramid Network: Pyramid networks are commonly used to differentiate objects at various scales. The

Functional Pyramid Network (FPN) uses the CNNs' unique multiscale pyramid order to highlight pyramids.

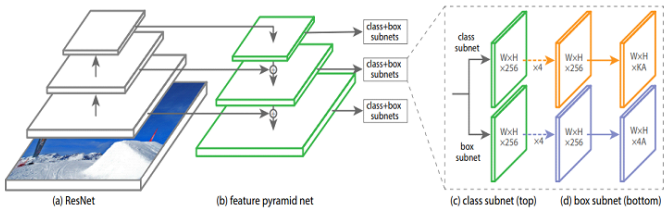


Fig. 4.

To generate a rich, multiscale convolutional inclusion pyramid (a), the one-stage RetinaNet network design leverages the Feature Pyramid Network (FPN) backbone in addition to feedforward ResNet engineering (b). RetinaNet joins the two subnetworks using this backbone. One is for anchor box characterization (c), and the other is for anchor box return to the ground truth object box (d). This job focuses on the original important accidental task that bridges the accuracy gap between single-stage locators and state-of-the-art two-stage identifiers such as Faster RCNN with FPN during execution because to the intended simplicity of organisational structure.

Focal Loss: Central Loss is a cross-entropy misfortune augmentation that helps to reduce overall misfortune for all around arranged models while putting greater emphasis on challenging, misclassified models.

The central issue allows us to create very accurate thick article indicators given a huge amount of simple basic models.

$$FL(P_t) = -\alpha_t(1 - P_t)^{\gamma}\, log(P_t) \tag{2}$$

IV. Experimental Results

A. Implementation

The deep learning neural community has the benefits of faster computation, higher accuracy, and easier operation. As a result, the deep studying technique has been widely used in item detection. Here we Implemented an green Methodology for Detection, Tracking, that is the hybrid deep RetinaNet-classifier for the research. The center concept in the back of deep studying is that complete characteristic representations may be successfully found out with the deep architectures which can be composed of stacked layers of trainable non-linear operations. However, due to the variety of video body content, it's far difficult to examine powerful characteristic representations at once from snap shots for Detection.

A new method to characteristic discount wherein characteristic choice and extraction and classifier schooling are executed concurrently the usage of a genetic set of rules. The genetic set of rules optimizes a characteristic weight vector used to scale the character capabilities within side the unique sample vectors. A overlaying vector is likewise hired for simultaneous choice of a characteristic subset. We appoint this approach in aggregate with the RetinaNet, and examine the consequences with classical characteristic choice and extraction techniques.

1) OBJECT STATE ESTIMATION WITH KALMAN FILTER: Our proposed monitoring set of rules is designed in Kalman clear out framework, on this paper, Kalman clear out goals to enhance the capacity of monitoring transferring item with a changeable speed, lowering the quest location and dealing with transient occlusion is a good optimization approach. Inspired with the aid of using nature, it's far designed for locating the best answer with the aid of using repeated choice and crossover operations. Genetic set of rules operates o people similar to genes within side the path of evolution, and every person represents a probable answer. In every generation, all people are evaluated with the aid of using a health feature to advantage affordable evolution. Genetic algorithm is simple, applicable for parallel processing, and has a excessive robustness in looking international optimum solution due to the fact it could take care of multi-mode distribution. In this paper, we use genetic set of rules to find the object, that is, looking the candidate location which is the maximum just like the goal model.

2) Classification: CNN In order for the device to pick out gadgets accurately, we want to teach it. We followed the CNN version to do so. The first layer is the enter layer in which the complete dataset(in batches) is fed into the community version. The technique of function extraction starts from the second one layer till it reaches the ultimate photo withinside the ultimate batch. A distinct function is extracted in every layer as the development of the neural community progresses. For example, first it extracts a point, then a line after which a curve. All the capabilities are mixed on the end (fully-linked layer) ensuing in an output layer.

3) Functional: First, an photo is taken and YOLO set of rules is applied. In our example, the photo is split as grids of 3x3 matrixes. We can divide the photo into any variety grids, relying at the complexity of the photo. Once the photo is split, every grid undergoes class and localization of the item. The objectiveness or the self belief rating of every grid is determined. If no right item can be found within the grid, the objectiveness and bounding field fee of the grid can be 0; if an item can be found within the grid, the objectiveness can be 1 and the bounding field fee can be the determined item's matching bounding values. The bounding field prediction is defined as follows. Also, Anchor packing containers are used to boom the accuracy of item detection which additionally defined beneath in detail.

YOLO, in a unique glance, takes the whole photo and predicts for those packing containers the bounding field co-ordinates and sophistication probabilities. YOLO's best gain is its first-rate pace, it is extraordinarily fast, and it may deal with forty five frames according to second [1]. Amongst the

3 variations of YOLO, version-three is quickest and extra correct in phrases of detecting small objects. The proposed set of rules, YOLO version-three includes overall 106 layers [10]. The structure is made from three wonderful layer forms. Firstly, the residual layer that is fashioned while activation is effortlessly forwarded to a deeper layer within side the neural network. In a residual setup, outputs of layer 1 are delivered to the outputs of layer 2. Second is the detection layer which plays detection at three distinctive scales or stages. Size of the grids is expanded for detection. Third is the up-sampling layer which will increase the spatial decision of an photo. Here photo is up sampled earlier than its miles scaled. Also, concatenation operation is used, to concatenate the outputs of preceding layer to the existing layer.

From the result it is clear that the stationary objects are properly detected as shown in Fig:5. the accuracy of the proposed scheme gives 99% accuracy for single object and 95% accuracy for multiple object detection.

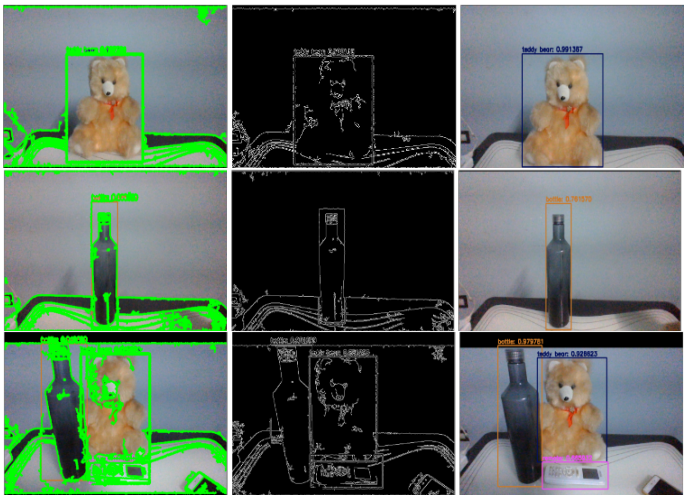


Fig. 5. Different stationary object detection results

V. Conclusion

In this paper a stable object detection scheme is proposed using RetinaNet and Kalman filter to detect stationary objects. Using of Kalman filter achieves the prediction of the out of focus regions. Also the use of RetinaNet achieves the accuracy level of stationary object detection. It is observed that the proposed scheme gives stable results for single and multiple object detection.

References

[1] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.

[2] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan, "Visual object detection with deformable part models," *Communications of the ACM*, vol. 56, no. 9, pp. 97–105, 2013.

[3] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," pp. 580–587, 2014.

[4] K. Wang, Y. Dong, H. Bai, Y. Zhao, and K. Hu, "Use fast r-cnn and cascade structure for face detection," pp. 1–4, 2016.

[5] J. Dai, Y. Li, K. He, and J. Sun, "R-fcn: Object detection via region-based fully convolutional networks," *Advances in neural information processing systems*, vol. 29, 2016.

[6] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," pp. 779–788, 2016.

[7] M. Ulrich, C. Steger, and A. Baumgartner, "Real-time object recognition using a modified generalized hough transform," *Pattern Recognition*, vol. 36, no. 11, pp. 2557–2570, 2003.

[8] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," pp. 2980–2988, 2017.

[9] M. Afif, R. Ayachi, Y. Said, E. Pissaloux, and M. Atri, "An evaluation of retinanet on indoor object detection for blind and visually impaired persons assistance navigation," *Neural Processing Letters*, vol. 51, no. 3, pp. 2265–2279, 2020.

[10] Y. Li, A. Dua, and F. Ren, "Light-weight retinanet for object detection on edge devices," pp. 1–6, 2020.

[11] S. J. Julier and J. K. Uhlmann, "New extension of the kalman filter to nonlinear systems," vol. 3068, pp. 182–194, 1997.

[12] S. Haykin, "Kalman filtering and neural networks," vol. 47, 2004.

[13] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," pp. 21–37, 2016.

[14] J. Sun, J. Wang, and T. Yeh, "Video understanding: from video classification to captioning," *Computer vision and pattern recognition. Stanford University*, pp. 1–9, 2017.

[15] M. Chihaoui, A. Elkefi, W. Bellil, and C. B. Amar, "Detection and tracking of the moving objects in a video sequence by geodesic active contour," pp. 212–215, 2016.

[16] B. Tian, L. Li, Y. Qu, and L. Yan, "Video object detection for tractability with deep learning method," pp. 397–401, 2017.

[17] S. Safie, A. A. Samah, G. Sulong, H. A. Majid, R. Muhammad, and H. Hasan, "Block matching algorithm for moving object detection in video forensic," pp. 1–5, 2017.