International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 08 | August - 2023 SJIF Rating: 8.176 ISSN: 2582-3930

A Novel Architecture for Object Detection and Tracking

J.Thilagavathy
Assistant professor

R.Mercy
PG Student

Applied Electronics and Engineering

Grace College Of Engineering, Tuticorin

ABSTRACT

The Convolutional Neural Networks (CNN) is used for the image recognition in the field of computer vision. Convolutional single–stage object detectors have been able to efficiently detect object of various sizes using a feature pyramid network. However, because they adopt a too simple manner of aggregating feature maps, they cannot avoid performance degradation due to information loss. The global information extractor is designed so that each feature vector that can reflect the information of the entire image through a proposed deep neural network with performance enhancement. Feature fusionmodel along with SE each channel so as to extract all necessary features support to improve the accuracy of the detector. Object detection and tracking is one of the most common and demanding tasks that system need to perform in order to detect meaningful events and activity to automatically comment and retrieve video content. The reason object detection and tracking are grouped is that object detection can be considered the basis of object tracking, and everyone needs to choose the right features and train for effective classification.

1. Introduction:

Artificial intelligence is an important branch of computer application, and the study of artificial intelligence aims of make machines think and humanly react to the outside world so that they can perform complex tasks that are beyond the reach of humans Important research direction in the field of image. processing, image recognition is the most basic task of computer vision and the basis of various other vision tasks.The single-stage model extracts feature maps at several stage of CNN.

As an important research direction in the field of image processing, image recognition is the most basic task of computer vision and the basis of various other vision tasks. For example, in the field of autonomous driving, the use of cameras to distinguish obstacles such as trees, animals, and pedestrians in front of the driver can effectively help drivers avoid hazards .

The image data set is an important driving for cein the development of image recognition, which contains 15 million images and more than 20,000 categories. The ImageNet dataset contains data from 15 million images and more than 20,000 categories. The emergence of the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) based on this dataset has facilitated the rapid development of deep learning algorithms in the field of image recognition, achieving results that far exceed those of general image recognition algorithms based on manual feature extraction, and the emergence of several algorithms such as Alex Net, Res Net, Inception Net, and other classic deep learning-based recognition algorithms. At the last edition of the competition in 2017, ImageNet announced the end of a large vision challenge based

on this dataset and will hence forth work on solving unsolved vision challenges.

In this competition, SE Net took the top spot in this image recognition with a top-5 error rate of 2.25, which has far surpassed human ability and shows the maturity of image recognition technology

2. Literature Survey

Object Detection Using Improved Bi-Directional Feature Pyramid Network:

Tran Ngoc Quang, Seunghyun Lee and Byung Cheol Song

Conventional single-stage object detectors have been able to efficiently detect objects of various sizes using a feature pyramid network. However, because they adopt a too simple manner of aggregating feature maps, they cannot avoid performance degradation due to information loss. To solve this problem, this paper proposes a new framework for single-stage object detection.However, it is true that the processing speed of the proposed method still does not reach SOTA. Therefore, further improving the computational efficiency is our future research task.

Image Classification Using Convolutional Neural Networks:

Deepika Jaswal, Sowmya.V, K.P.Soman

—Deep Learning has emerged as a new area in machine learning and is applied to a number of signal and image applications.The main purpose of the work presented in this paper, is to apply the concept of a Deep Learning algorithm namely, Convolutional neural networks (CNN) in image classification. The algorithm is tested on various standard datasets, like remote sensing data of aerial images (UC Merced Land Use Dataset) and scene images from SUN database. The performance of the algorithm is evaluated based on the quality metric known as Mean Squared Error (MSE) and classification accuracy.

Deep Residual Learning for Image Recognition:

Kaiming He Xiangyu Zhang Shaoqing Ren Jian Sun Microsoft Research Deeper neural networks are more difficult to train. We present a residual learning framework to ease the training of networks that are substantially deeper than those used previously. We explicitly reformulate the layers as learning residual functions with reference to the layer inputs, instead of learning unreferenced functions.

We provide comprehensive empirical evidence showing that these residual networks are easier to optimize, and can gain accuracy from considerably increased depth. On the ImageNet dataset we evaluate residual nets with a depth of up to 152 layers—8× deeper than VGG nets [40] but still having lower complexity.Our method has good generalization performance on other recognition tasks

Multiple Object Detection and Tracking :

Anbarasi MP Robotics and Automation Engineering PSG College of Technology,
Coimbatore Deep Learning for Multiple Object Tracking, summarizes and analyzes deep learning based multi-object tracking methods [2]. Methods fall into three categories: extendeddescription with deep capabilities**,** embedding deep networks, and building end-to-end deep networks. It then examines then deep network structure in these ways and describes the use and training of these networks for multi-object tracking problems. A performance improvement in the detection and tracking of multiple objects in real-time and online applications, and is concerned with surveillance and video traffic systems. Asignificant improvement in terms of tracking performance could be achieved with an accuracy of approximately 65%. Discovery - based Multi-Object Tracking Method, describes how to track multiple objects. The total number of objects are unidentified and will modify during tracking.

Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks:

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun

State-of-the-art object detection networks depend on region proposal algorithms to hypothesize object locations. Advances like SPP net [1] and Fast R-CNN [2] have reduced the running time of these detection networks, exposing region proposal computation as a bottleneck. In this work, we introduce a Region Proposal Network (RPN) that shares full-image convolutional features with the detection network, thus enabling early cost-free region proposals. An RPN is a fully convolutional network that simultaneously predicts object bounds and object ness scores at each position.

We have presented RPNs for efficient and accurate region proposal generation. By sharing convolutional A score threshold of 0.6 is used to display these images. For each image, one colour represents one object category in that image.

Very Deep Convolutional Networks For Large-Scale Image Recognition:

Karen Simonyan∗& Andrew Zisserman+ Visual Geometry Group, Department of Engineering Science, University of Oxford {karen,az}@robots.ox.ac.uk

In this work we investigate the effect of the convolutional network depth on its accuracy in the large-scale image recognition setting. Our main contribution is a thorough evaluation of networks of increasing depth using an architecture with very small (3×3) convolution filters, which shows that a significant improvement on the prior-art configurations can be achieved by pushing the depth to 16–19 weight layers. Our models generalise well to a wide range of tasks and datasets, matching or outperforming more complex recognition pipelines built around less deep image representations.

3. CNN ANALYSIS

3.1 NERUAL NETWORK:

A neural network is a network or circuit of neurons, or in a modern sense, an artificial neural network, composed of artificial neurons or nodes. Thus a neural network is either a biological neural network, made up of real biological neurons, or an artificial neural network, for solving artificial intelligence (AI) problem. The connections of the biological neuron are modelled as weights.

A positive weight reflects an excitatory connection, while negative values mean inhibitory connections. All inputs are modified by a weight and summed. This activity is referred as a linear combination. Finally, an activation function controls the amplitude of the output. For example, an acceptable range of output is usually between 0 and 1, or it could be - 1 and 1. These artificial networks may be used for predictive modelling , adaptive control and applications where they can be trained via a dataset. Self-learning resulting from experience can occur within networks, which can derive conclusions from a complex and seemingly unrelated set of information.

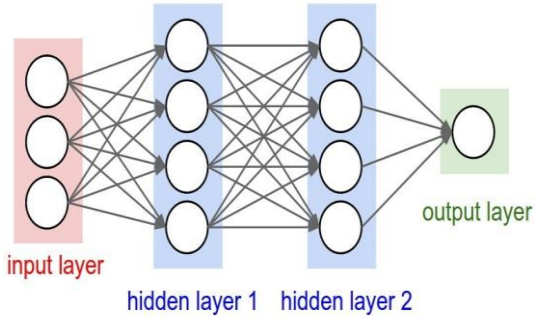

Fig.1 A simple neural network

Fig 1 is A deep neural network (DNN) is an artificial neural network (ANN) with multiple layers between the input and output layers. The DNN finds the correct mathematical manipulation to turn the input into the output,

whether it be a linear relationship or a non-linear relationship.

3.2 CONVOLUTIONAL NEURAL NETWORKS:

Convolutional Neural Networks are very similar to ordinary Neural Networks from the previous chapter: they are made up of neurons that have learnable weights and biases. Each neuron receives some inputs, performs a dot product and optionally follows it with nonlinearity.

Convolutional Neural Networks (CNNs) are analogous to traditional ANNs in that they are comprised of neurons that self-optimise through learning. Each neuron will still receive an input and perform a operation (such as a scalar product followed by a non-linear function) - the basis of countless ANNs. From the input raw image vectors to the final output of the class score, the entire of the network will still express a singleperceptive score function (the weight). The last layer will contain loss functions associated with the classes, and all of the regular tips and tricks developed for traditional ANNs still apply.

The only notable difference between CNNs and traditional ANNs is that CNNs are primarily used in the field of pattern recognition within images. This allows us to encode image-specific features into the architecture, making the network more suited for image-focused tasks - whilst further reducing the parameters required to set up the model. One of the largest limitations of traditional forms of ANN is that they tend to struggle with the computational complexity required to compute image data. Common machine learning benchmarking datasets such as the MNIST database of handwritten digits are suitable for most forms of ANN, due to its relatively small image dimensionality of just 28× 28.

With this dataset a single neuron in the first hidden layer willcontain 784 weights (28×28×1 where 1 bare in mind that MNIST is normalised to just black and white values), which is manageable for most forms of ANN.

If you consider a more substantial coloured image input of 64 × 64, the number of weights on just a single neuron of the first layer increases substantially to 12, 288. Also take into account that to deal with this scale of input, the network will also need to be a lot larger than one used to classify colour-normalised MNIST digits, then you will understand the drawbacks of using such models.

3.3 CNN ARCHITECTURE:

CNNs are feedforward networks in that information flow takes place in one direction only, from their inputs to their outputs. Just as artificial neural networks (ANN) are biologically inspired, so are CNNs. The visual cortex in the brain, which consists of alternating layers of simple and complex cells (Hubel & Wiesel, 1959, 1962), motivates their architecture.

CNN architectures come in several variations; however, in general, they consist of convolutional and pooling (or subsampling) layers, which are grouped into modules. Either one or more fully connected layers, as in a standard feedforward neural network, follow these modules. Modules are often stacked on top of each other to form a deep model. It illustrates typical CNN architecture for a toy image classification task. An image is input directly to the network, and this is followed by several stages of convolution and pooling. Thereafter, representations from these operations feed one or more fully connected layers.

Finally, the last fully connected layer outputs the class label. Despite this being the most popular base architecture found in the literature, several architecture changes have been proposed in recent years with the objective of improving image classification accuracy or reducing computation costs. Although for the remainder of this section, we merely fleetingly introduce ssstandard CNN architecture.

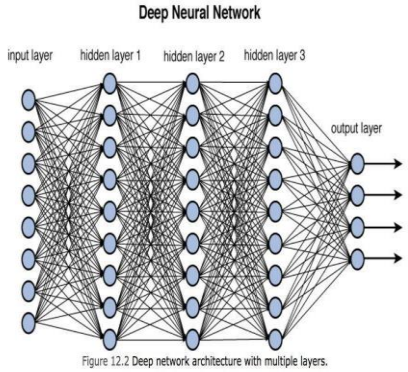

Fig 2 Deep Neural Network

3.4 OVERALL ARCHITECTURE:

CNNs are comprised of three types of layers. These are convolutional layers, pooling layers and fullyconnected layers. When these layers are stacked, a CNN architecture has been formed. A simplified CNN architecture for MNIST classification is illustrated 0 9 convolution w/ReLu pooling output fully-connected w/ReLu fully-connected architecture, comprised of just five layers .

The basic functionality of the example CNN above can be broken down into four key areas. 1. As found in other forms of ANN, the input layer will hold the pixel values of the image. 2. The convolutional layer will determine the output of neurons of which are connected to local regions of the input through the calculation of the scalar product between their weights and the region connected to the input volume. The rectified linear unit (commonly shortened to ReLu) aims to apply an 'elementwise' activation function such as sigmoid to the output of the activation produced by the previous layer. 3. The pooling layer will then simply perform downsampling along the spatial dimensionality of the given input, further reducing the number of parameters within that activation. 4. The fully-connected layers will then perform the same duties found in standard ANNs and attempt to produce class scores from the activations, to be used for classification. It is also suggested that ReLu may be used between these layers, as to improve performance. Through this simple method of transformation, CNNs are able to transform the original input layer by layer using convolutional and downsampling techniques to produce class scores for classification and regression purposes.

3.5 CONVOLUTIONAL LAYERS:

The convolutional layers serve as feature extractors, and thus they learn the feature representations of their input images. The neurons in the convolutional layers are arranged into feature maps.

Each neuron in a feature map has a receptive field, which is connected to a neighbourhood of neurons in the previous layer via a set of trainable weights, sometimes referred to as a filter bank. Inputs are convolved with the learned weights in order to compute a new feature map, and the convolved results are sent through a nonlinear activation function.

All neurons within a feature map have weights that are constrained to be equal; however, different feature maps within the same convolutional layer have different weights so that several features can be extracted at each location. As the name implies, the convolutional layer plays a vital role in how CNNs operate. The layers parameters focus around the use of learnable kernels.

These kernels are usually small in spatial dimensionality, but spreads along the entirety of the depth of the input. When the data hits a convolutional layer, the layer convolves each filter across the spatial dimensionality of the input to produce a 2D activation map. These activation maps can be visualised.As we glide through the input, the scalar product is calculated for each value in that kernel. From this the network will learn kernels that 'fire' when they see a specific feature at a given spatial position ofthe input. These are commonly known as activations.

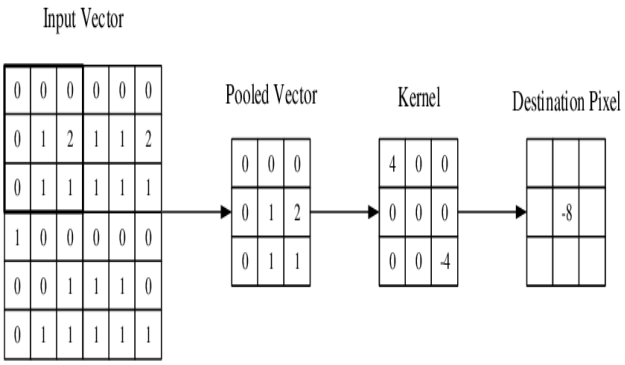

Fig: 3 Visual representation of a convolutional layer

The centre element of the kernel is placed over the input vector, of which is then calculated and replaced with a weighted sum of itself and any nearby pixels. Every kernel will have a corresponding activation map, of which will be stacked along the depth dimension to form the full output volume from the convolutional layer.

3.6 Pooling Layers :

The purpose of the pooling layers is to reduce the spatial resolution of the feature maps and thus achieve spatial invariance to input distortions and translations. Initially, it was common practice to use average pooling aggregation layers to propagate the average of all the input values, of a small neighbourhood of an image to the next layer. However, in more recent models, max pooling aggregation layers propagate the maximum value within a receptive field to the next layer.

Pooling layers aim to gradually reduce the dimensionality of the representation, and thus further reduce the number of parameters and the computational complexity of the model. The pooling layer operates over each activation map in the input, and scales its dimensionality using the "MAX" function. In most CNNs, these come in the form of max-pooling layers with kernels of a dimensionality of 2 × 2 applied with a stride of 2 along the spatial dimensions of the input. This scales the activation map down to 25% of the original size - whilst maintaining the depth volume to its standard size.

3.7 Fully Connected Layers:

Several convolutional and pooling layers are usually stacked on top of each other to extract more abstract feature representations in moving through the network. The full connected layers that follow these layers interpret these feature representations and perform the function of high-level reasoning. For classification problems, it is standard to use the soft max operator on top of a DCNN. While early success was enjoyed by using radial basis functions (RBFs), as the classifier on top of the convolutional towers found that replacing the soft max operator with a support vector machine (SVM) leads to improved classification accuracy.

The fully-connected layer contains neurons of which are directly connected to the neurons in the two adjacent layers, without being connected to any layers within them. This is analogous to way

that neurons are arranged in traditional forms of ANN.

Convolutional Neural Networks differ to other forms of Artifical Neural Network in that instead of focusing on the entirety of the problem domain, knowledge about the specific type of input is exploited. This in turn allows for a much simpler network architecture to be set up.

3.8 Training:

CNNs and ANN in general use learning algorithms to adjust their free parameters in order to attain the desired network output. The most common algorithm used for this purpose is backpropagation. Backpropagation computes the gradient of an objective function to determine how to adjust a network's parameters in order to minimize errors that affect performance. A commonly experienced problem with training CNNs, and in particular DCNNs, is overfitting, which is poor performance on a held-out test set after the network is trained on a small or even large training set. This affects the model's ability to generalize on unseen data and is a major challenge for DCNNs that can be assuaged by regularization.

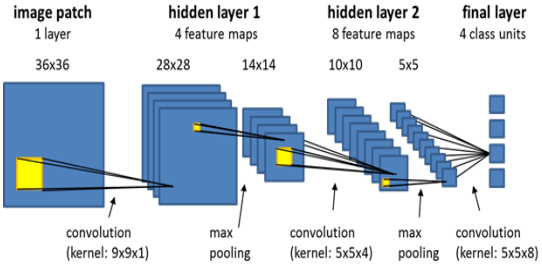

Fig 4 Convolutional layers

4. PROPOSED SYSTEM

4.1 METHODS:

The convolutional neuralnetworks will inevitably require the high accuracy of CNN models. In practical applications, the target scenes are often more or less different from the image set at the time of training, and how to make the model maintain good adaptability and accuracy in complex environments is the issue focused on in this study.

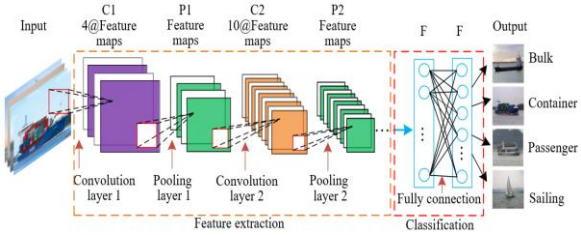

Fig 5. sample methods

4.2 BLOCK DIAGRAM:

Object Detection and Tracking

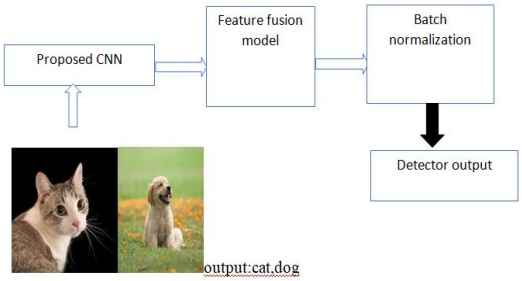

Fig 6 Block diagram object detection and Tracking

Feature fusion model:

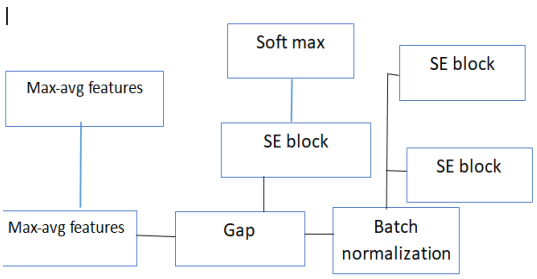

Fig 7 Feature Fusion Modal

Features map:

Generated by applying filters or feature detectors to the input image or the feature map output of the prior layer.

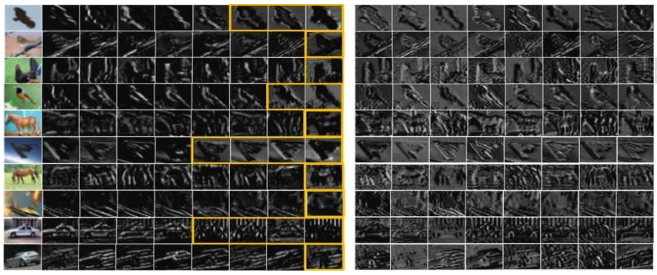

Fig 8: feature map

5. SOFTWARE REQUIREMENT

5.1 MATLAB:

MATLAB is a proprietary multi-paradigm programming language and numeric computing environment developed by MathWorks. MATLAB allows matrix manipulations, plotting of functions and data, implementation of algorithms, creation of user interfaces, and interfacing with programs written in other languages. The heart of MATLAB is the MATLAB language, a matrix-based language allowing the most natural expression of computational mathematics.

5.2 Features of MATLAB:

Following are the basic features of MATLAB:

- It is a high-level language for numerical computation, visualization and application development.
- It also provides an interactive environment for iterative exploration, design and problem solving.
- It provides vast library of mathematical functions for linear algebra, statistics, Fourier analysis, filtering, optimization, numerical integration and solving ordinary differential equations.
- It provides built-in graphics for visualizing data and tools for creating custom plots.
- MATLAB's programming interface gives development tools for improving code quality, maintainability,and maximizing performance.
- It provides tools for building applications with custom graphical interfaces.
- It provides functions for integrating MATLAB based algorithms with external applications and languages such as C, Java, .NET and Microsoft Excel.

5.3 Uses of MATLAB:

MATLAB is widely used as a computational tool in science and engineering encompassing the fields of physics, chemistry, math and all engineering streams. It is used in a range of applications including

- signal processing and Communications
- image and video Processing
- control systems
- test and measurement
- computational finance
- computational biology

6.Experimental results

Some real time Images of Human Face and Object are taken for Training and Testing. Our architecture shows 85.5% accuracy for Human Detection and 95.6% accuracy in Object detection.

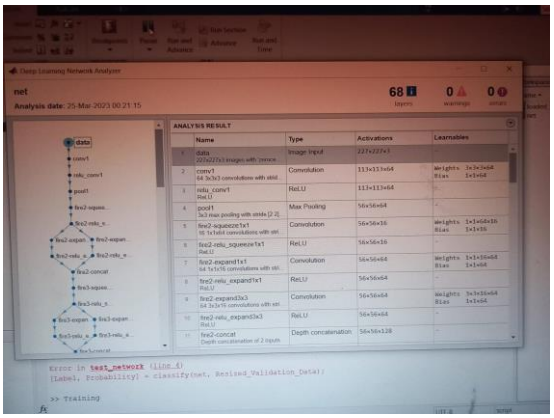

Fig 9 Network analysis

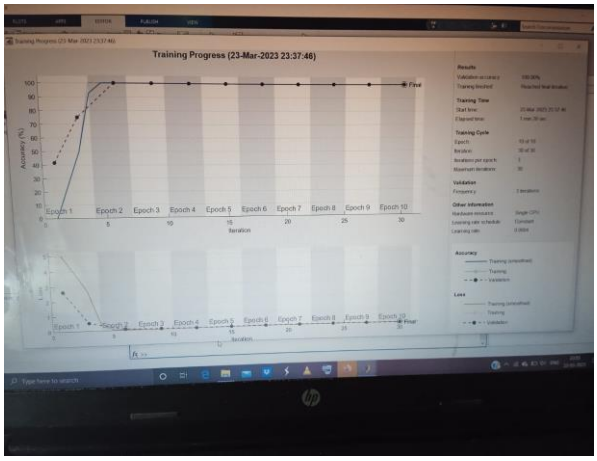

Fig 10 Training

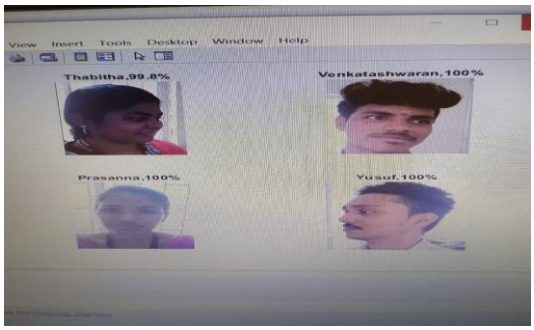

Fig 11 Human Face Identification

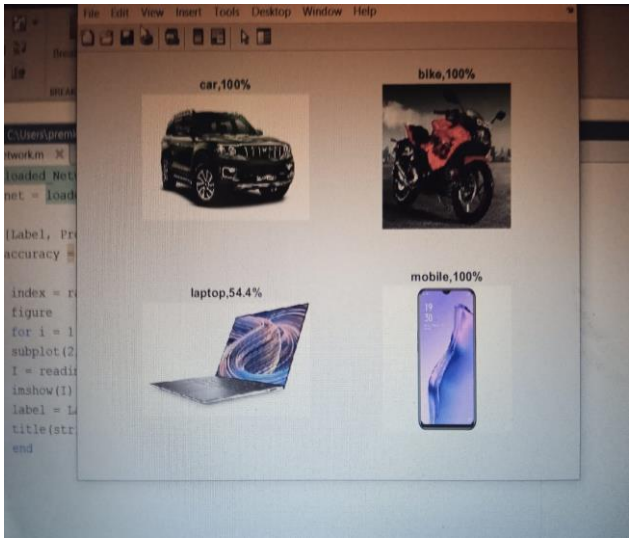

Fig 12 Object Identification

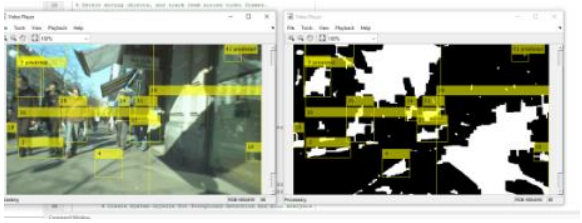

Fig 13 Tracking of Human and Object using Kalman Filter

8 CONCULSION

Image recognition using CNN base papers are problem time is very large to decrease the time and the quality of speed less to increase CNN. Object detector are only at the image for the consider but the papers consider same of the image to detection. Aiming at the problem of low accuracy of scene recognition, an image detection algorithm based on saliency is designed to eliminate the background information in the image, highlight the effective features of the image, and reduce the interference caused by illumination change and perspective change in the process of recognition. This methods to detect and tracking the videos at tracks in the objects frame at create the predict images. Videos to propose the tracking at the objects in low time. Fast tracking at the motion based multiple objects. In long-duration videos, hard and easy scenes are not separate problems and spatial target
distribution can be quite non- homogenous.

REFERENCES

1. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In Advances in Neural Information Processing Systems; Neural Information Processing Systems Foundation Inc.: San Diego, CA, USA, 2015; pp. 91–99.

2. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.

3. Liu, W.; Anguelov, D.; Erhan, D.; Sze gedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In European Conference on Computer Vision; Springer: Cham, Switzerland, 2016; pp. 21–37.

4. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.

5. Zhao,Q.;Sheng,T.;Wang,Y.;Tang,Z.;Chen,Y.;Cai,L. ;Ling,H.M2det: A single-shot object detector basedon multi-level feature pyramid network. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 9259–9266.

6. Tan, M.; Pang, R.; Le, Q.V. Efficient det: Scalable and efficient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10781–10790.

7. 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 (ECCV), 2014.

8. R. Girshick, “Fast R-CNN,” in IEEE International Conference on Computer Vision (ICCV), 2015.