e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:05/May-2022 Impact Factor- 6.752 www.irjmets.com

OBJECT DETECTION AND HUMAN ACTIVITY RECOGNITION

Harbhajan Sharma[*1], Robin Kumar[*2], Sharvan Sharma[*3], Mr. Sanjeev Raghav[*4]

[*1,2,3]Dept. Of Computer Science & Engineering, HMR Institute Of Technology & Management, Delhi, India.

[*4]Assistant Professor, Dept. Of Computer Science & Engineering, HMR Institute Of Technology & Management, Delhi, India.

ABSTRACT

The field of computer vision known as real-time object detection is vast, dynamic, and sophisticated. Image Localization is used when there is only one object to differentiate in an image, while Object Detection is used when there are several objects in an image. The most often used strategies for contemporary deep learning models to fulfill various tasks on embedded devices are mobile networks and binary neural networks. We present a method for identifying an item based on the MobileNet deep learning pre-prepared model for Single Shot Multibox Detector in this study (SSD). This method is used to detect objects in a video stream in real time and to broadcast webcams. Then, to determine what is in the video stream, we utilize an object detection module. We combine the MobileNet and SSD frameworks to construct a quick and efficient deep learning-based item identification solution to finish the module. Activity recognition has been a significant field of research in recent decades. Humans can recognise activities in their environment based on a range of observations. These observations are used in a wide range of applications, including video surveillance, health care, gesture detection, energy conservation, and fall detection systems. Sensor-based techniques like accelerometers and gyroscopes have been investigated, as well as their advantages and disadvantages. Sensors can be employed in a variety of ways in a properly regulated environment. This work takes a step-by-step approach to developing a human activity recognition system. In convolutional neural networks for image classification, the Resnet-34 model is used to provide shortcut connections that solve the vanishing gradient problem. The model has been successfully trained and tested, providing a good result, by recognising over 400 human activities. Finally, some unresolved topics are highlighted, which should be researched more in the future.

Keywords: Open CV,SSD (Single Shot Detector), Activity, Dataset.

I. INTRODUCTION

The field of computer vision known as real-time object detection is vast, dynamic, and sophisticated. Image Localization is used when there is only one object to differentiate in an image, while Object Detection is used when there are several objects in an image. The most often used strategies for contemporary deep learning models to fulfill various tasks on embedded devices are mobile networks and binary neural networks. We present a method for identifying an item based on the MobileNet deep learning pre-prepared model for Single Shot Multibox Detector in this study (SSD). This method is used to detect objects in a video stream in real time and to broadcast webcams. Then, to determine what is in the video stream, we utilize an object detection module. We combine the MobileNet and SSD frameworks to construct a quick and efficient deep learning-based item identification solution to finish the module.

Human activity recognition has received a lot of interest in recent decades. The data acquired by activity tracking can be used for a variety of objectives, including safe driving, lowering crime rates, and initiating appropriate medical treatment activities, to name a few. Activity detection may also benefit elderly folks by making their lives easier and simpler. Individuals have the ability to monitor and recognise other humans' actions. communication. Machines, on the other hand, must learn before they can recognise activities. The activity recognizer can detect walking, applause, reading, hand washing, and a range of other movements. Human-to-human interactions include handshakes, hugs, and other activities involving two individuals; meanwhile, human-to-object interactions, which involve one human and one item, include reading a book or newspaper. Some of these exercises are simple while others are really difficult. To make complex activities more understandable, they can be broken down into smaller steps. Accelerometers, sensors, photographs, and video frames are just some of the ways data can be collected. When collecting data with sensors, people must

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:05/May-2022 Impact Factor- 6.752 www.irjmets.com

wear many sensors in various locations on their body. The acquired data must be processed in a number of ways. After the raw data has been segmented, different features are extracted. This method may be problematic in the case of sensors. In this paper, the use of a deep neural network aids in the extraction of more meaningful data. Following that, classification methods are applied to these features. The classification model is based on a training dataset and is used to recognise activities. Hidden Markov Models (HMMs), Support Vector Machine Classifiers, and Feed-forward neural networks are a few classification algorithms.

In this research, the Resnet-34 algorithm is employed to develop an intelligent human activity recognition system. One of the most popular image classification architectures is Resnet. It contains shortcut connections that allow a signal to skip a layer and proceed directly to the next in the sequence. It is made up of two convolutional layers, each with batch normalization and a rectified linear unit (ReLu). As the network converges further, the accuracy of simple neural networks, which essentially stack layers, diminishes. As a result, the main advantage of using the Resnet model is that it overcomes the problem of vanishing gradients. This model is easy to tweak and has a smaller training error than others.

II. METHODOLOGY

1. Object Detection

a. Mobilenet-SSD

In our recommended approach, we leverage the MobileNet- SSD architecture. One of the reasons we chose this design is because it gives excellent object detection accuracy while being quicker than competing designs like YOLO, as proven in the study. This is especially true when using low-powered computer devices, such as those in our system, to identify an object in real time. MobileNet-SSD achieves a quicker detection time by addressing the model with 8-bit integers rather than 32-bit floats. The model took a 300 by 300 pixel image as input and produced the bounding box position as well as detection confidences (ranging from 0 to 1) for each recognised object. A detection confidence level of 0.5 was employed to assess if the detected item was genuine.

b. OpenCV (Open-Source computer vision)

OpenCV is a computer vision programming library that focuses on real-time computer vision. OpenCV is a free and open-source computer vision framework that may be used to analyze CCTV footage, videos, and images. It's an excellent programme for employment in image processing and computer vision. OpenCV is a C++ library with approximately 2,500 high-performance algorithms. We don't want to start from scratch when developing computer vision apps; instead, we may utilize this library to focus on real-world difficulties. The OpenCV method cv2.VideoCapture() may read video. We may access the camera by passing 0 as a function parameter.

2. Human Activity Recognition

Training and recognition are the two most important elements in the implementation process. The method is stochastic gradient descent. The training data is used to generate training samples. Next, for the production of training samples, a temporal place in the video is chosen. A sixteen-frame clip is then created around the designated spot. If the video is less than sixteen frames, begin looping it as many times as necessary. Following that, a geographical position and spatial scale are chosen based on the requirements. Crop four corners to 112 X 112 pixels using the corner cropping approach. The training parameter includes a 0.001 weight decay. After the validation loss reaches a saturation point, the learning rate is reduced to 0.01. For recognition, each frame is looped from 0 to sample duration. Reading the video gets a frame and returns true if the picture is successfully obtained. After that, the frame is scaled down to 400 pixels and added to the collection of frames. After then, the frame is tweaked once again to ensure that all of the photographs are the same size. After the frame array has been completely filled, the blob will be generated. A blob of input frames is created by the network and transferred across it. The image must first be preprocessed in order to provide the correct prediction. By removing the mean value and scaling it by a factor of 0.1, the image is preprocessed.

This decreases the image's sensitivity to backdrop and lighting conditions. The deep neural network module in OpenCV is used to do all of this. All of the images in this blob were produced to have the same spatial dimensions. The algorithm then does a forward pass, grabbing a label with the correct prediction value.

The Kinetic dataset is used to train the model. About 650,000 edited videos with a duration of about 10 seconds are included in this dataset, which is split into 400 categories of human activities.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:05/May-2022 Impact Factor- 6.752 www.irjmets.com

When videos are scaled, the aspect ratio remains unchanged. The dataset includes applause, archery, bartending, tears, handshakes, and a variety of other behaviors. This collection includes frames that are connected to target action. Because it is free of noise and disconnected frames, this dataset is ideal for training. In all, there are about 580000, 30000, and 40000 training, validation, and test sets. When the Resnet-34 model is trained on a kinetic dataset, there is no overfitting. The results of this experiment might have a big influence on future computer vision developments.

III. MODELING AND ANALYSIS

The suggested system would quickly and efficiently recognise objects in real time utilizing the Mobilenet-SSD idea. We'll create a Python script for object detection with a deep neural network using OpenCV 3.4.

The following is how the system works:

The simplified MobileNet Architecture, which generates light-weight deep neural networks using depth-wise separable convolutions, will be used to deliver input via real-time video from a camera or webcam. The input video is divided into frames and sent to the MobileNet layers. [4] The amount of pixel intensity in the bright zone is subtracted from the amount of pixel intensity in the dark region to arrive at each feature value. All of the image's accessible sizes and regions are used to compute these components. There may be many irrelevant traits in an image, but just a few important ones that can be used to detect the item.

The objective of the MobileNet layers is to turn the pixels in the input image into highlights that define the image's content. The MobileNet-SSD model is then used to calculate the bounding boxes and corresponding class (label) of objects. The only thing left to do now is present or show the output.

THE PROPOSED SYSTEM ARCHITECTURE IN DIAGRAM:

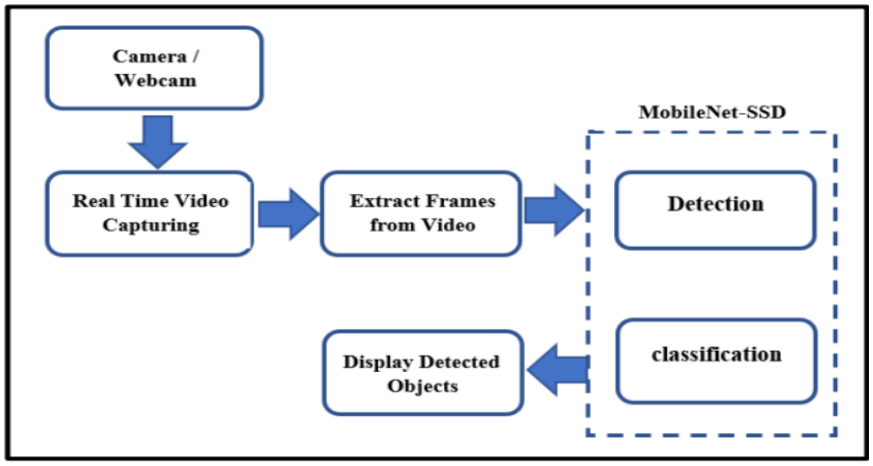

Figure 1: The Proposed System Architecture In Diagram.

IV. RESULTS AND DISCUSSION

1. Object Detection

Figure 2: Horse Detection.

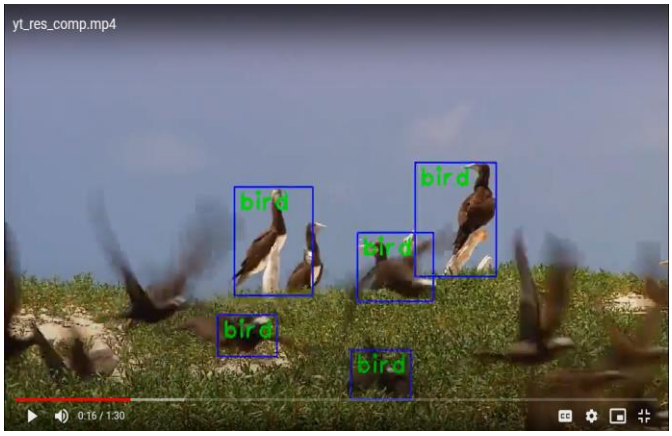

Figure 3: Birds Detection.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:05/May-2022 Impact Factor- 6.752 www.irjmets.com

2. Human Activity And Recognition

The validation loss on the kinetic dataset is greater than the training loss, showing that the Resnet-34 model does not overfit the kinetic dataset.

The findings imply that the kinetic dataset may be used to train resnet-34 without overfitting. An average accuracy of 71.2 is achieved on the kinetic validation set over Top-1 and Top-5. As a result, kinetics can build a deep 3D CNN from the ground up. UCF50 is an action recognition data set with 50 action categories, consisting of realistic videos taken from youtube. This data set is an extension of YouTube Action data set (UCF11) which has 11 action categories. Our data set is very challenging due to large variations in camera motion, object appearance and pose, object scale, viewpoint, cluttered background, illumination conditions, etc.

In TensorFlow, the Media Sequence library provides a comprehensive set of capabilities for storing data. Sequence Examples. Sequence Examples are efficient to employ with TensorFlow and provide matching semantics to most video jobs. The sequence semantics support a variable number of annotations per frame, which is required for jobs like video object detection but difficult to implement in TensorFlow. Examples. Media Sequences objective is to make dealing with Sequence Examples easier and to automate common preparation activities. MediaPipe Pose is a machine learning solution for high-fidelity body pose tracking that infers 33 3D landmarks and a background segmentation mask on the complete body from RGB video frames, based on our BlazePose research, which also drives the ML Kit Pose Detection API. Current state-of-the-art approaches for inference rely mostly on sophisticated desktop environments, whereas our solution runs in real time on the latest mobile phones, desktops/laptops, Python, and even the web.

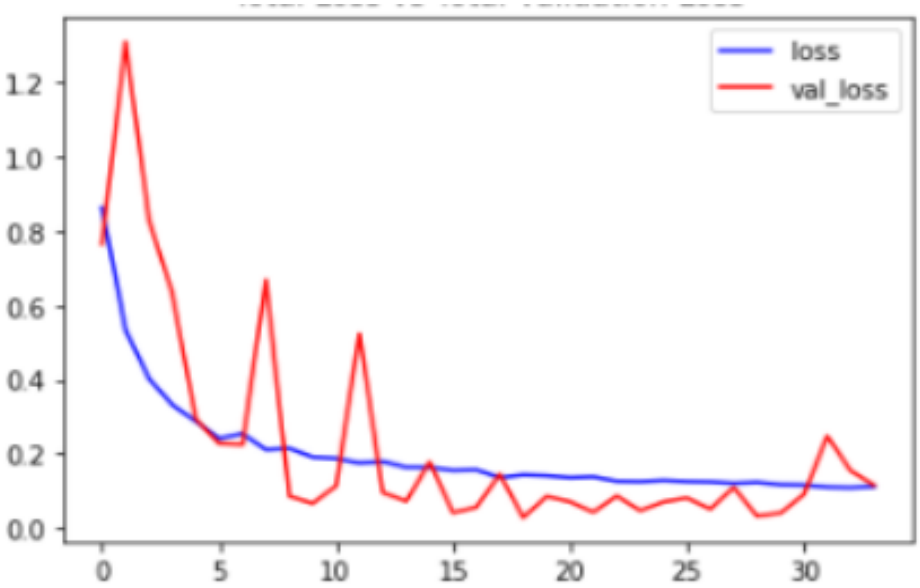

Figure 4: Total Loss Vs Validation Loss.

Figure 5: Horse Detection Detection.

Figure 6: TaiChi Activity Detection.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:05/May-2022 Impact Factor- 6.752 www.irjmets.com

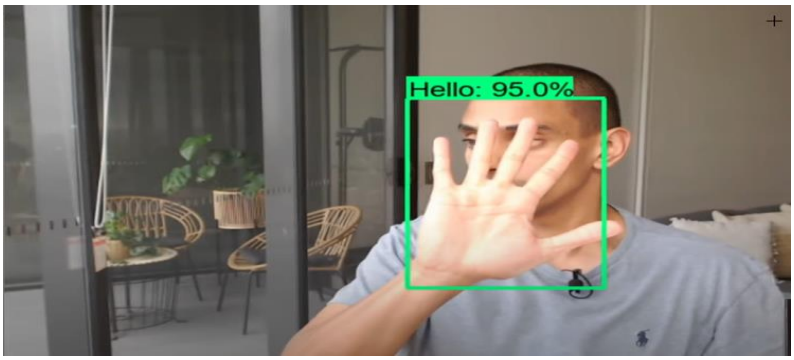

Figure 7: Hello Activity Detection.

V. CONCLUSION

This document explains the process of activity recognition as well as alternate activity recognition methods. Image recognition has become a crucial research area in the field of computer vision. Football, eating, dancing, and other activities are examples of actions. Deep learning models can learn from simple to complicated data because of their layer-by-layer structure. All of this is feasible because to the capabilities of today's computers and databases. Deep learning is always coming up with new ways to solve problems. Activity identification is useful in a variety of applications, including smart homes and ambient living. In order to sustain a global state of law and peace, it is becoming increasingly vital to recognise behaviors such as fighting in public places and vandalism.

In terms of activity recognition, there has been a lot of progress. Other challenges, such as recognising intricate and concurrent activities, exist. Things like walking while listening to music or singing while dancing are examples of simultaneous activity. These patterns of behavior become puzzling and difficult to discern.

Many more research are being carried out in order to fully address these challenges. Sensor-based systems have additional challenges, such as placing sensors on various parts of the human body to detect activity directly. Users find it difficult to wear sensors in watches, clothes, bracelets, and other products. External sensors are strategically positioned across the environment in various locations. Because GPS receivers are confined to the outside, sensors can only be utilized in specific locations. Sensors must be put in every door and piece of equipment in a smart house. Setting up and maintaining such a large network takes a long time. With the aid of cameras, these sensors might be replaced. The Resnet-34 model is utilized to track the training and recognition process in this article. The procedure for performing the assignment is well outlined. The model has been rigorously tested to guarantee that it produces the expected results. Because it tracks many layers of skip connections, Resnet-34 produces the best results. These skip weights might be named using a weight matrix.

The Kinetic dataset is used to produce nearly 400 different human behaviors, with a decent degree of realism. The video clips are of excellent quality and were taken from YouTube. A single clip may contain many actions in some circumstances. If two or more activities occur at the same time, such as "texting" while "walking" or "eating" while "chatting," only one of the classes will be labeled, not both. To distinguish, some activities, such as playing various sorts of musical instruments, need a greater concentration on the thing.

ACKNOWLEDGEMENTS

The authors would like to thank the Department of Computer Science and Engineering at HMRITM for their assistance with this project. We are grateful to our mentor, Mr. Sanjeev Raghav, for giving us such a wonderful and challenging opportunity, and we thank our college's HOD for providing us with all of the necessary assistance and encouragement.

VI. REFERENCES

[1] Harshal Honmote, Pranav Katta, Shreyas Gadekar, Madhavi Kulkarni, “Real Time Object Detection and Recognition using MobileNet-SSD with OpenCV”, International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181, Vol. 11 Issue 01, January-2022

[2] Akansha Abrol, Anisha Sharma, Kritika Karnic, Raju Ranjan, “Human Activity Recognition using Resnet-34 Model”, International Journal of Recent Technology and Engineering (IJRTE), ISSN: 2277-3878

(Online), Volume-10 Issue-1, May 2021

[3] N. Albukhary and Y.M. Mustafah 2017 IOP Conf. Ser.: Mater. Sci. Eng. **260** 012017.

[4] Andres Heredia and Gabriel Barros-Gavilanes," Video processing inside embedded devices using SSD-Mobilenet to count mobility actors," 978-1-7281-1614-3/19 ©2019 IEEE.

[5] G. Bradski and, A. Kaehler, "Learning OpenCV", OReilly Publications,2008.

[6] R. Huang, J. Pedoeem, and C. Chen, "YOLO-LITE: A Real-Time Object Detection Algorithm Optimized for Non-GPU Computers," in Proceedings - 2018 IEEE International Conference on Big Data, Big Data 2018.

[7] R. Huang, J. Pedoeem, and C. Chen, "YOLO-LITE: A Real-Time Object Detection Algorithm Optimized for Non-GPU Computers," in Proceedings - 2018 IEEE International Conference on Big Data, Big Data 2018.