International Conference on Advances in Materials, Computing and Communication Technologies
In Association withInternational Journal of Scientific Research in Science and Technology
Volume 9 | Issue 1 | Print ISSN: 2395-6011 | Online ISSN: 2395-602X (www.ijsrst.com)

Deep Learning Technique to Detect Object For Visually Impaired People Using YOLO V3 Framework Mechanism

Dr. S. Saraswathi, N. Subashi, M. Sneha, I. Amithapbatchan
Department of Information Technology, Pondicherry Engineering College, Puducherry, India

ABSTRACT

In this project, we recommended a technique called the multi-view object tracking (MVOT) system to resolve the multiple cameras monitor an area from different angles. Videos recorded by the cameras contain complementary information and fusing the knowledge embedded in the videos facilitates the development of a robust and accurate system. Those task of cameras that have different settings, we propose a correspondence Yolo V3 algorithm that maps each segmented group of objects in one view to the corresponding group in another view. We call these corresponding groups matched blob clusters, each of which enables knowledge to be shared between cameras. It follows that we present a two-pass regression framework for multi-view objects.

Keywords : Object Detection, convolutional Neural Network, Object tracking, moving object detection, you only Look Once.

I. INTRODUCTION

The term digital image refers to processing of a two dimensional picture by a digital computer. In a broader context, it implies digital processing of any two dimensional data. A digital image is an array of real or complex numbers represented by a finite number of bits. An image given in the form of a transparency, slide, photograph or an X-ray is first digitized and stored as a matrix of binary digits in computer memory. This digitized image can then be processed and/or displayed on a high-resolution television monitor. For display, the image is stored in a rapid-access buffer memory, which refreshes the monitor at a rate of 25 frames per second to produce a visually continuous display.

II. LITERATURE SURVEY

Y. WU, J. LIM, AND M.-H. YANG, ``ONLINE OBJECT TRACKING: A BENCHMARK," IN PROC. IEEE COMPUT. VIS. PATTERN RECOGNIT., JUN. 2019, PP. 24112418.

Visual object tracking is the process of tracking an arbitrary object in a video, where the bounding box of the object is given in the first frame. Siamese network-based visual object tracking approaches have recently received considerable attention for their high speed and superior performance. However, for scale and angle estimation, Siamese trackers require multiple search regions, which increases the computation time, thereby decreasing the real-time

tracking performance. This paper proposes a one-shot Siamese network, named Siam-OS, for fast and efficient visual object tracking. Siam-OS uses only a single search region and estimates the scale and angle of the target bounding box. This significantly reduces the number of computations required for the deep convolutional feature extraction, and thus increases the tracking speed. The experimental results with Visual Object Tracking (VOT) benchmarks show the effectiveness of the proposed Siam-OS in terms of the accuracy, robustness, expected average overlap, and speed.

M. DANELLJAN AND G. HÄGER, F. KHAN, AND M. FELSBERG, ``ACCURATE SCALE ESTIMATION FOR ROBUST VISUAL TRACKING," IN PROC. BRIT. MACH. VIS. CONF. NOTTINGHAM, U.K.: BMVA PRESS, SEP. 2018.

Object detection in streaming images is a major step in different detection-based applications, such as object tracking, action recognition, robot navigation, and visual surveillance applications. In most cases, image quality is noisy and biased, and as a result, the data distributions are disturbed and imbalanced. Most object detection approaches, such as the faster region-based convolutional neural network (RCNN), single shot multibox detector with 300×300 inputs (SSD300), and you only look once version 2 (YOLOv2), rely on simple sampling without considering distortions and noise under real-world changing environments, despite poor object labeling. In this paper, we propose an incremental active semi-supervised learning (IASSL) technology for unseen object detection. It combines batch-based active learning (AL) and bin-based semi-supervised learning (SSL) to leverage the strong points of AL's exploration and SSL's exploitation capabilities. A collaborative sampling method is also adopted to measure the uncertainty and diversity of AL and the confidence in SSL. Batch-based AL allows us to select more informative, condent, and representative samples with low cost. Bin-based SSL divides streaming image samples into several bins, and each bin repeatedly transfers the discriminative knowledge of convolutional neural network deep learning to the next bin until the performance criterion is reached. The IASSL can overcome noisy and biased labels in unknown, cluttered data distributions. We obtain superior performance, compared with the state-of-the-art technologies, such as Faster RCNN, SSD300, and YOLOv2.

J. HAN, R. QUAN, D. ZHANG, AND F. NIE, ``ROBUST OBJECT CO-SEGMENTATION USING BACKGROUND PRIOR," IEEE TRANS. IMAGE PROCESS., VOL. 27, NO. 4, PP. 16391651, APR. 2018.

Given a set of images that contain objects from a common category, object co-segmentation aims at automatically discovering and segmenting such common objects from each image. During the past few years, object co-segmentation has received great attention in the computer vision community. However, the existing approaches are usually designed with misleading assumptions, unscalable priors, or subjective computational models, which do not have sufficient robustness for dealing with complex and unconstrained real-world image contents. This paper proposes a novel two-stage co-segmentation framework, mainly for addressing the robustness issue. In the proposed framework, we first introduce the concept of union background and use it to improve the robustness for suppressing the image backgrounds contained by the given image groups. Then, we also weaken the requirement for the strong prior knowledge by using the background prior instead. This can improve the robustness when scaling up for the unconstrained image contents. Based on the weak background prior, we propose a novel MR-SGS model, i.e., manifold ranking with the self-learned graph structure, which can infer suitable graph structures in a data-driven manner rather than building the fixed graph structure relying on the subjective design. Such capacity is critical for further improving the robustness in inferring the foreground/background probability of each image pixel. Comprehensive experiments and comparisons

with other state-of-the-art approaches can demonstrate the effectiveness of the proposed work.

J. KWON AND K. M. LEE, ``TRACKING OF A NON-RIGID OBJECT VIA PATCH-BASED DYNAMIC APPEARANCE MODELING AND ADAPTIVE BASIN HOPPING MONTE CARLO SAMPLING," IN PROC. IEEE CONF. COMPUT. VIS. PATTERN RECOGNIT., JUN. 2019, PP. 12081215.

We propose a novel tracking algorithm for the target of which geometric appearance changes drastically over time. To track it, we present a local patch-based appearance model and provide an efficient scheme to evolve the topology between local patches by on-line update. In the process of on-line update, the robustness of each patch in the model is estimated by a new method of measurement which analyzes the landscape of local mode of the patch. This patch can be moved, deleted or newly added, which gives more flexibility to the model. Additionally, we introduce the Basin Hopping Monte Carlo (BHMC) sampling method to our tracking problem to reduce the computational complexity and deal with the problem of getting trapped in local minima. The BHMC method makes it possible for our appearance model to consist of enough numbers of patches. Since BHMC uses the same local optimizer that is used in the appearance modeling, it can be efficiently integrated into our tracking framework. Experimental results show that our approach tracks the object whose geometric appearance is drastically changing, accurately and robustly.

B. J. LEE ET AL., ``PERCEPTION-ACTION-LEARNING SYSTEM FOR MOBILE SOCIALSERVICE ROBOTS USING DEEP LEARNING," IN PROC. ASSOC. ADVANCEMENT ARTIF.INTELL. CONF. ARTIF. INTELL. (AAAI), FEB. 2018.

We introduce a robust integrated perception-action-learning system for mobile social-service robots. The state-of-the-art deep learning techniques were incorporated into each module which significantly improves the performance in solving social service tasks. However, our system is yet to fulfill every individual's expectations on performance and processing speed, our demonstration highlights the importance of research on not only the individual elements, but the integration of each modules for developing a more human-like, idealistic robot to assist humans in the future.

X. LI, Q. LIU, N. FAN, Z. HE, AND H. WANG, ``HIERARCHICAL SPATIAL-AWARE SIAMESE NETWORK FOR THERMAL INFRARED OBJECT TRACKING," KNOWL.-BASED SYST., VOL. 166, PP. 7181, FEB. 2019.

Constructing a robust appearance model of the visual object is a crucial task for visual object tracking. Recently, more and more studies combine spatial feature with a temporal feature to improve the tracking performance. These methods successfully apply the features from spatial and temporal to address the problem for tracking. This paper presents a novel method for visual object tracking based on spatiotemporal feature combined with correlation filters. In this paper, the visual features of a target object are extracted from a spatialtemporal residual network (STResNet) appearance model with two sub-networks. The STResNet appearance model learns separately spatial feature and temporal feature, respectively, so that we can effectively utilize spatial context around the surrounding of the target object in each frame and the temporal relationship between successive frames to rene the appearance representation of the target object. Finally, our spatiotemporal fusion feature from STResNet appearance model is incorporated into the correlation filter for robust visual object tracking. The experimental results show that our method achieves similar or better performance compared with the other tracking methods based on convolutional neural networks or correlation filter.

A. LUKEZIC, T. VOJIR, L. C. ZAJC, J. MATAS, AND M. KRISTAN, ``DISCRIMINATIVE CORRELATION LTER WITH CHANNEL AND SPATIAL RELIABILITY," IN PROC. IEEE CONF. COMPUT.

VIS. PATTERN RECOGNIT., JUN. 2017, PP. 63096318.

Tracking-by-detection for visual object tracking is the most popular and successful framework at present. It treats the tracking problem as a classification task and learns information about the target from each tracking result online. Accurate model learning of the classifier requires numerous positive samples. However, it is difcult to obtain numerous positive training samples at the beginning of visual tracking. In this paper, we propose a novel comparative object similarity learning method to strengthen the training samples set. The core of our approach is that the comparative object similarity information between the candidate objects is taken into account when training classifiers. In addition, the classifier model is updated with the image information of the target to be predicted by further exploring the temporal context between successive image frames. According to the Bayesian inference theorem, the tracking results, which are estimated from the posterior probability distribution of target, are more accurate. We implement two versions of the proposed tracker with the representations from both conventional hand-crafted and deep convolution neural networks based features to validate the effectiveness of the algorithm. The quantitative and qualitative experimental results demonstrate that the proposed method performs superiorly against several state-of-the-art algorithms on large-scale challenging benchmark datasets.

J. CHOI, H. J. CHANG, S. YUN, T. FISCHER, Y. DEMIRIS, AND J. Y. CHOI, ``ATTENTIONAL CORRELATION LTER NETWORK FOR ADAPTIVE VISUAL TRACKING," IN PROC. IEEE CONF. COMPUT. VIS. PATTERN RECOGNIT., VOL. 2, JUL. 2017, PP. 48284837.

While moving ahead with the object detection technology, especially deep neural networks, many related tasks, such as medical application and industrial automation, have achieved great success. However, the detection of objects with multiple aspect ratios and scales is still a key problem. This paper proposes a top-down and bottom-up feature pyramid network (TDBU-FPN), which combines multi-scale feature representation and anchor generation at multiple aspect ratios. First, in order to build the multi-scale feature map, this paper puts a number of fully convolutional layers after the backbone. Second, to link neighboring feature maps, top-down and bottom-up flows are adopted to introduce context information via top-down flow and supplement sub-original information via bottom-up flow. The top-down flow refers to the deconvolution procedure, and the bottom-up flow refers to the pooling procedure. Third, the problem of adapting different object aspect ratios is tackled via many anchor shapes with different aspect ratios on each multi-scale feature map. The proposed method is evaluated on the pattern analysis, statistical modeling and computational learning visual object classes (PASCAL VOC) dataset and reaches an accuracy of 79%, which exhibits a 1.8% improvement with a detection speed of 23 fps.

M. WANG, Y. LIU, AND Z. HUANG, ``LARGE MARGIN OBJECT TRACKING WITH CIRCULANT FEATURE MAPS," IN PROC. IEEE CONF. COMPUT. VIS. PATTERN RECOGNIT., HONOLULU, HI, USA, JUL.2017, PP. 48004808.

Structured output support vector machine (SVM) based tracking algorithms have shown favorable performance recently. Nonetheless, the time-consuming candidate sampling and complex optimization limit their real-time applications. In this paper, we propose a novel large margin object tracking method which absorbs the strong discriminative ability from structured output SVM and speeds up by the correlation filter algorithm significantly. Secondly, a multimodal target detection technique is proposed to improve the target localization precision and prevent model drift introduced by similar objects or background noise. Thirdly, we exploit the feedback from high-confidence tracking results to avoid the model

corruption problem. We implement two versions of the proposed tracker with the representations from both conventional hand-crafted and deep convolution neural networks (CNNs) based features to validate the strong compatibility of the algorithm. The experimental results demonstrate that the proposed tracker performs superiorly against several state-of-the-art algorithms on the challenging benchmark sequences while runs at speed in excess of 80 frames per second.

REN S, HE K, GIRSHICK R, ET AL. FASTER R-CNN: TOWARDS REAL-TIME OBJECT DETECTION WITH REGION PROPOSAL NETWORKS. PROC. OF THE ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS, 2015: 91 – 99.

State-of-the-art object detection networks depend on region proposal algorithms to hypothesize object locations. Advances like SPPnet [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 nearly cost-free region proposals. An RPN is a fully convolutional network that simultaneously predicts object bounds and objectness scores at each position. The RPN is trained end-to-end to generate high-quality region proposals, which are used by Fast R-CNN for detection. We further merge RPN and Fast R-CNN into a single network by sharing their convolutional features-using the recently popular terminology of neural networks with 'attention' mechanisms, the RPN component tells the unified network where to look. For the very deep VGG-16 model [3], our detection system has a frame rate of 5 fps (including all steps) on a GPU, while achieving state-of-the-art object detection accuracy on PASCAL VOC 2007, 2012, and MS COCO datasets with only 300 proposals per image. In ILSVRC and COCO 2015 competitions, Faster R-CNN and RPN are the foundations of the 1st-place winning entries in several tracks. Code has been made publicly available.

the literature survey is done from an existing system which has some limitations in Intra camera visual are not much effective and accurate when adapting the different aspects of object counting and those limitations are considered in our project

III. PROPOSED METHODOLOGY

It has been often seen that in an uncontrolled condition only some regions of the whole image are affected by facial changes due to variations of pose, illumination, expression, etc. The conventional appearance-based global feature extraction methods are usually applied on the whole face image. As a result, these methods are not suitable to cope with above mentioned local facial changes. Therefore, to improve the robustness of a face recognition system, it is necessary to consider local features from these parts of the face region along with the global features in the feature extraction process.

Our methodology involves two main modules. The first stage involves acquiring image from the web camera and converting it into text document using Optical Character Recognition (OCR). The second stage involves natural language processing and digital signal processing for converting the text into speech using Text to Speech synthesizer (TTS).

Steps involved in our methodology

1. Image acquisition by the web camera
2. Loading the image into the axial panel of the created Graphical User Interface (GUI)
3. Pre-processing of the image (RGB to gray image, contrast adjustment, adaptive threshold)
4. Converting pre-processed image into text document using OCR
5. Converting text document into speech using TTS.

Image analysis is the extraction of meaningful information from images; mainly from digital images by means of digital image processing techniques. Many important image analysis tools such as edge detectors and neural networks are inspired by human visual perception models. Computer image analysis

largely contains the fields of computer or machine vision, and medical imaging, and makes heavy use of pattern recognition, digital geometry, and signal processing.

A. YOLO V3

YOLO V3 consider classification as one of the most dynamic research and application areas. Yolo V3 is the branch of Artificial Intelligence (AI). The neural network was trained by Yolo V3 algorithm. The different combinations of functions and its effect while using Yolo V3 as a classifier is studied and the correctness of these functions are analyzed for various kinds of datasets. The YoloV3 can be used as a highly successful tool for dataset classification with suitable combination of training, learning and transfer functions. When the maximum likelihood method was compared with COCO method, the Yolo v3 was more accurate than maximum likelihood method. A high predictive ability with stable and well functioningYolo V3 is possible. It proves to be more effective than other classification algorithms.

B. COCO Method

Though the layers are colloquially referred to as convolutions, this is only by convention. Mathematically, it is technically a sliding dot product or cross-correlation. This has significance for the indices in the matrix, in that it affects how weight is determined at a specific index point.

Convolutional networks may include local or global pooling layers to streamline the underlying computation. Pooling layers reduce the dimensions of the data by combining the outputs of neuron clusters at one layer into a single neuron in the next layer. Local pooling combines small clusters, typically 2 x 2. Global pooling acts on all the neurons of the convolutional layer. In addition, pooling may compute a max or an average. Max pooling uses the maximum value from each of a cluster of neurons at the prior layer. Average pooling uses the average value from each of a cluster of neurons at the prior layer.

C. TEXT TO SPEECH SYNTHESIS (TTS)

Speech synthesis is the artificial production of human speech. A computer system used for this purpose is called a speech synthesizer, and can be implemented in software or hardware. A text-to-speech (TTS) system converts normal language text into speech; other systems render symbolic linguistic representations like phonetic transcriptions into speech. 16 Synthesized speech can be created by concatenating pieces of recorded speech that are stored in a database. Systems differ in the size of the stored speech units; a system that stores phones or diaphones provides the largest output range, but may lack clarity. For specific usage domains, the storage of entire words or sentences allows for high-quality output. The quality of a speech synthesizer is judged by its similarity to the human voice, and by its ability to be understood. An intelligible text-to-speech program allows people with visual impairments or reading disabilities to listen to written works on a home computer. Text-to-Speech (TTS) refers to the ability of computers to read text aloud. A TTS Engine converts written text to a phonemic representation, then converts the phonemic representation to waveforms that can be output as sound. TTS engines with different languages, dialects and specialized vocabularies are available through third party publishers.

TABLE I

S. No	Analytical Reasoning		
	Technique	Situation applicable	Benefits
1	Yolo Only Look Once Version3	Training learning and transfer functions. Real Time App. Face	Easy to implement and better detecting smaller objects. It is incredibly fast and accuracy(91-95%)

		Detection	process 45 frames per sec.
2	Text To Speech Synthesis	To Aid Visually impaired by offering a computer generated spoken voice that would read text to the user.	The storage of entire words or sentences allows for high-quality output. 16 Synthesized speech can be created by concatenating pieces of recorded speech that are stored in a database.

TABLE II

S. No	Comparison of Features with Existing Work			
	Feature	Method	Existing Work	Proposed Work
1	Variable number of Objects	Common Object on Context	CNN (convolutional Neural Network).	Dual Priorities: Object Classification and Localization
2	Limited Data	Common Object on Context	Crowd sourcing often produces image classification	Limited amount of annotated data currently and currently available for object detection.

D. Figure Captions

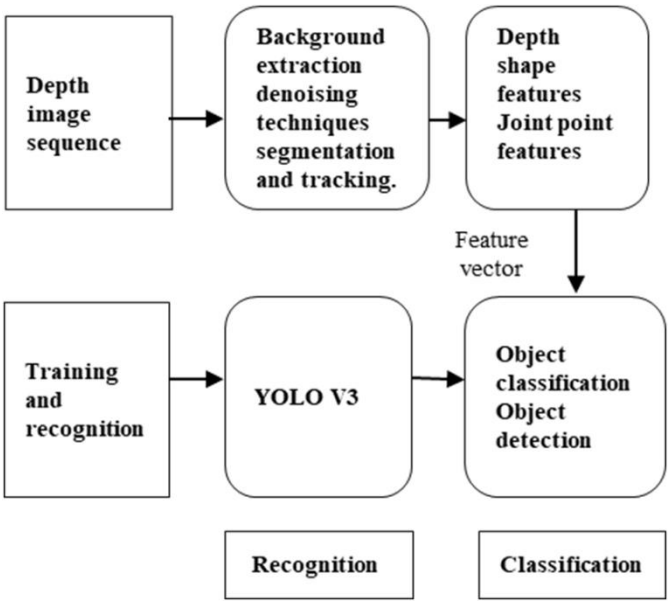


Fig. 1 Overall System Architecture

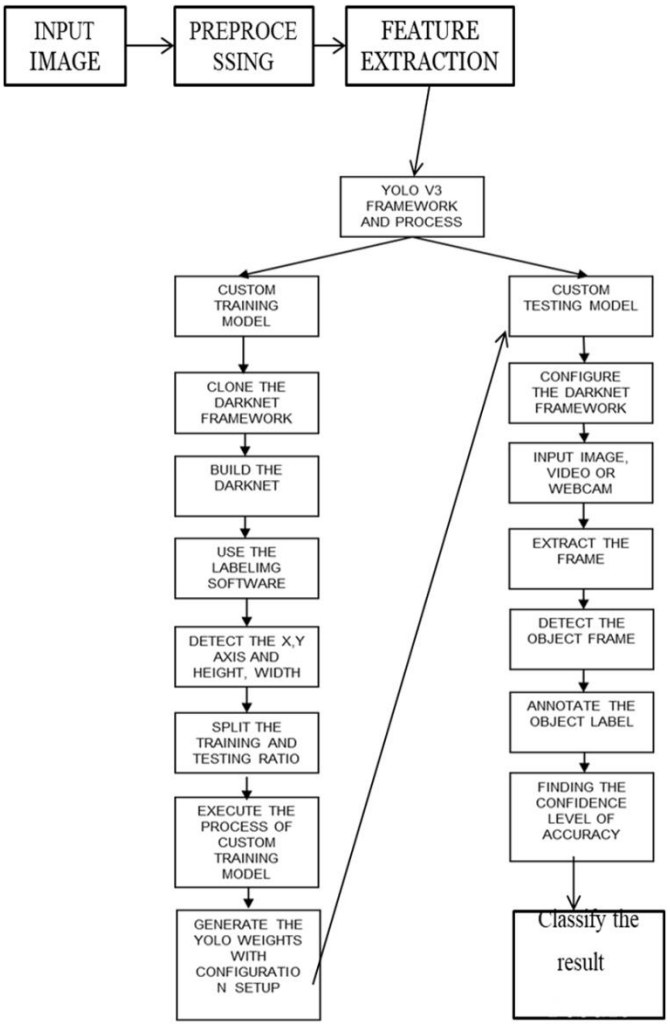


Fig. 2Yolo Object DetectionFlow Diagram

E. Algorithm with Pseudocode

YOLO stands for "You Only Look Once" and uses convolutional neural networks (CNN) for object detection.

1. When YOLO works it predicts classes' labels and detects locations of objects at the same time. That is why, YOLO can detect multiple objects in one image.
2. The name of the algorithm means that a single network just once is applied to whole image.
3. YOLO divides image into regions, predicts bounding boxes and probabilities for every such region.
4. YOLO also predicts confidence for every bounding box showing information that this particular bounding box actually includes object, and probability of included object in bounding box being a particular class.
5. Then, bounding boxes are filtered with technique called non-maximum suppression that excludes some of them if confidence is low or there is another bounding box for this region with higher confidence.
6. YOLO-3 is the latest version that uses successive 3x3 and 1x1 convolutional layers. In total it has 53 convolutional layers with architecture. Every layer is followed by batch normalization and Leaky ReLU activation

	Type	Filters	Size	Output
	Convolutional	32	3 × 3	256 × 256
	Convolutional	64	3 × 3 / 2	128 × 128
1×	Convolutional	32	1 × 1	
	Convolutional	64	3 × 3	
	Residual			128 × 128
	Convolutional	128	3 × 3 / 2	64 × 64
2×	Convolutional	64	1 × 1	
	Convolutional	128	3 × 3	
	Residual			64 × 64
	Convolutional	256	3 × 3 / 2	32 × 32
8×	Convolutional	128	1 × 1	
	Convolutional	256	3 × 3	
	Residual			32 × 32
	Convolutional	512	3 × 3 / 2	16 × 16
8×	Convolutional	256	1 × 1	
	Convolutional	512	3 × 3	
	Residual			16 × 16
	Convolutional	1024	3 × 3 / 2	8 × 8
4×	Convolutional	512	1 × 1	
	Convolutional	1024	3 × 3	
	Residual			8 × 8
	Avgpool		Global	
	Connected		1000	
	Softmax			

Fig. 3 Architecture of YOLO V3

batch=64–number of samples that will be processed in one batch

subdivisions=16–number of mini batches in one batch; GPU processes mini batch samples at once; the weights will be updated for batch samples, that is 1 iteration processes batch images

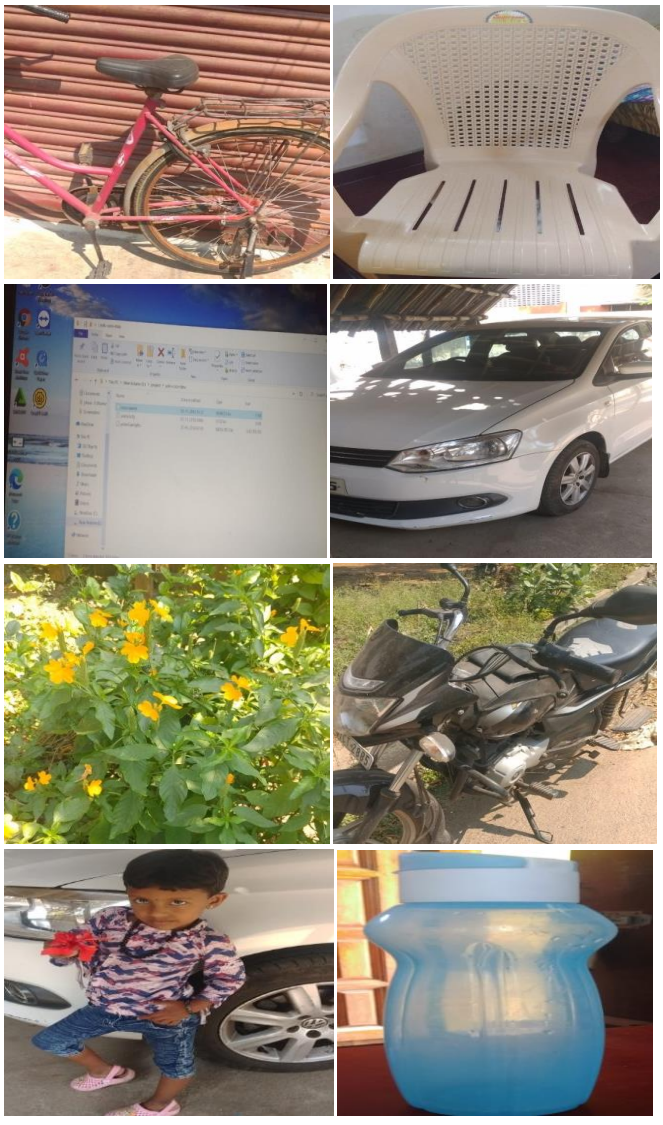
width=608–every image will be resized during training and testing to this number

height=608 – every image will be resized during training and testing to this number

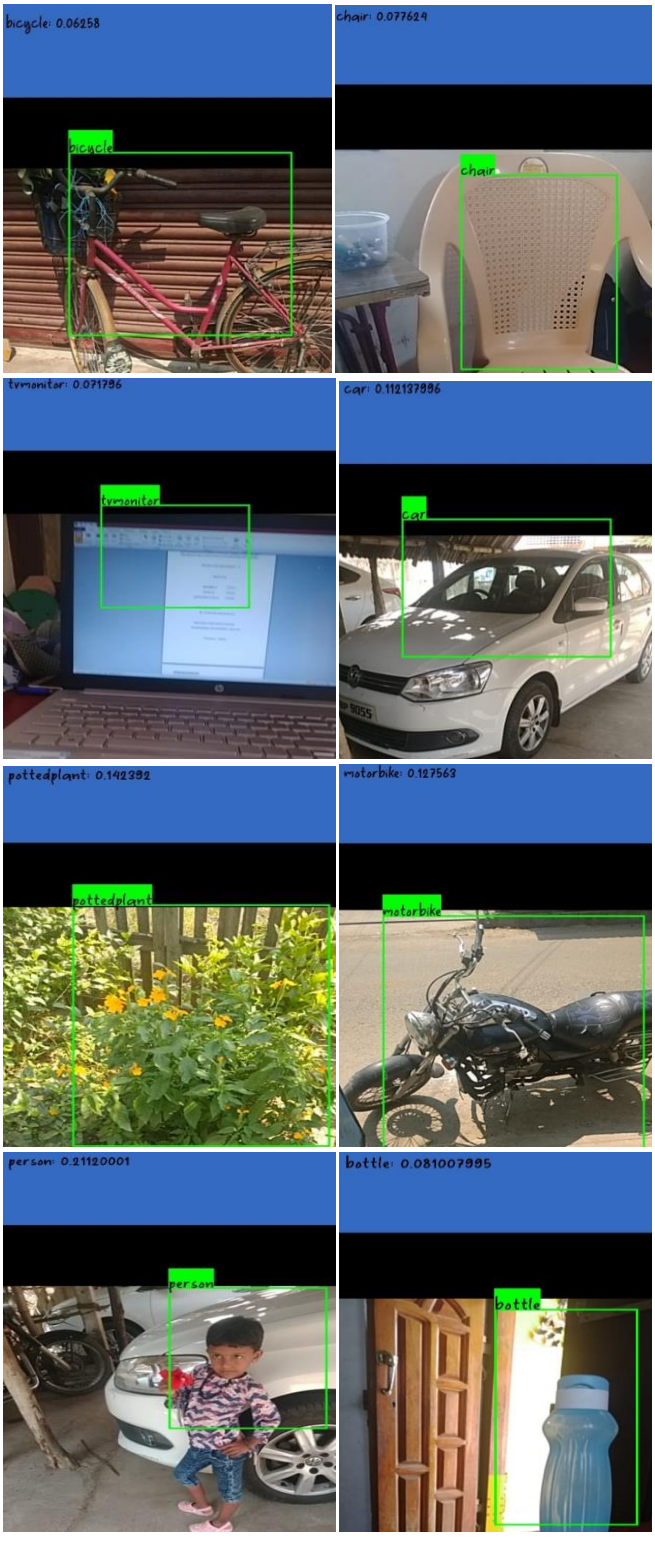
channels=3 – every image will be converted during training and testing to this number

IV. RESULT AND ANALYSIS

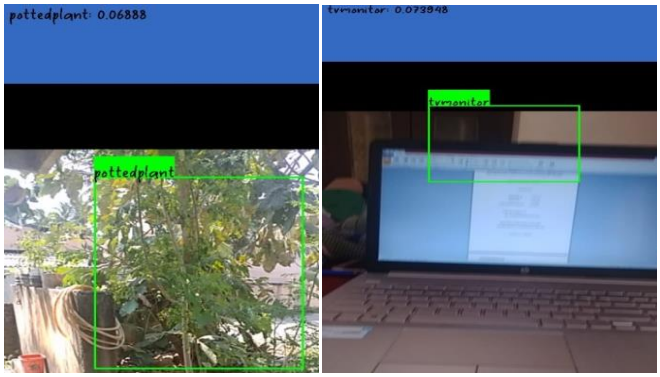
Input Specification



Output Specification

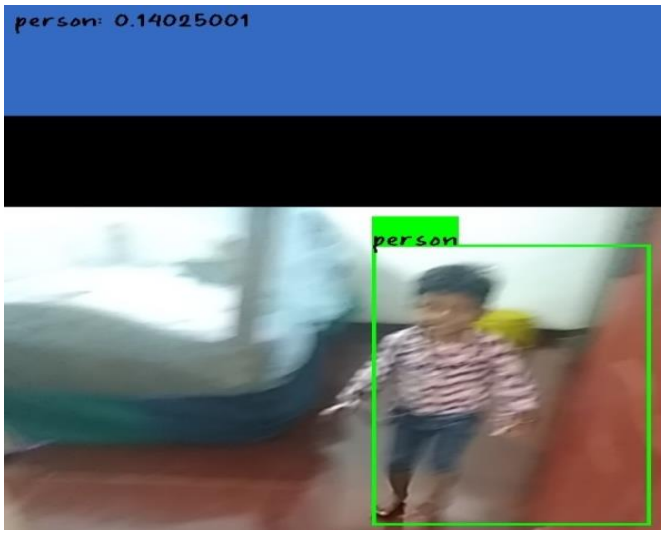


Blurred Output



Different types of Output

Moving Output



V. CONCLUSION

This projects shows that object counting and tracking are key activities in many mobile applications. The detection of moving objects uses a background subtraction algorithm based on Gaussian mixture models. Morphological operations are applied to the resulting foreground mask to eliminate noise. Finally, Gabor filter mechanism detects groups of connected pixels, which are likely to correspond to moving objects. The use of the YOLO v3 for tracking objects and focuses on three important features namely 1) Prediction of object's future location. 2) Reduction of noise introduced by inaccurate detections. 3) Facilitating the process of association of multiple objects to their tracks. This method is to detect and Counting the objects can be used to analyze in any platforms. Detection is also a first step prior to performing more sophisticated tasks such as tracking or categorization of objects by their type.

VI. REFERENCES

[1]. D.Velmurugan, M.S.Sonam, S.Umamaheswari, S.Parthasarathy, K.R.Arun[2016]. A Smart Reader for Visually Impaired People Using Raspberry PI. International Journal of Engineering Science and Computing IJESC Volume 6 Issue No. 3.S. M. Metev.

[2]. K NirmalaKumari, Meghana Reddy J [2016]. Image Text to Speech Conversion Using OCR Technique in Raspberry Pi. International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering Vol. 5, Issue 5, May 2016..,

[3]. SilvioFerreira, C´elineThillou, Bernard Gosselin. From Picture to Speech: An Innovative Application for Embedded Environment. FaculténPolytechnique de Mons, Laboratoire de Théorie des Circuits et Traitement du Signal Bˆatiment Multitel - Initialis, 1, avenue Copernic, 7000, Mons, Belgium

[4]. NagarajaL, Nagarjun R S, Nishanth M Anand, Nithin D, Veena S Murthy [2015]. Vision based Text Recognition using Raspberry Pi. International Journal of Computer Applications (0975 – 8887) National Conference on Power Systems & Industrial Automation

[5]. World Health Organisation. 10 facts about blindness and visual impairment 2015