International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

Enhancement of Real-Time Object Detection and Tracking in Collaborative Environment using AI and Mixed Reality

[1]Anurag Tiwari, [2]Nileshkumar Patel, [3]Shishir Kumar
[1]Research Scholar, Jaypee University of Engineering and Technology, Guna, MP, India
Email ID:- er.anurag907@gmail.com
Orcid ID:- 0009-0006-4455-3637
[2]Assistant Professor, Jaypee University of Engineering and Technology, Guna, MP, India
Email ID:- nilesh.juet@gmail.com
Orcid ID:- 0000-0001-6562-5982
[3]Professor, Babasaheb Bhimrao Ambedkar University (A Central University), Lucknow, UP, India
Email ID:- _dr.shishir@yahoo.com
Orcid ID:- 0000-0002-6850-653X

Abstract

The area of mixed reality has had rapid growth in recent years, with a notable rise in funding. This may be attributed to the rising recognition of the potential advantages associated with the integration of virtual information into the physical environment. The majority of contemporary mixed reality apps that rely on markers use algorithms for local feature identification and tracking. This study aims to enhance the accuracy of object recognition in complicated environment and enable real-time classification operations via the introduction of a unique detection approach known as the lightweight and efficient YOLOv4 model. In the present setting, Computational vision emerges as a very valuable and engaging manifestation of artificial intelligence (AI) that finds widespread application in many aspects of daily existence. The field of computer vision is dedicated to the development of advanced artificial intelligence and computer systems that aim to replace complex elements of the human environment. In recent times, deep neural networks have emerged as a crucial component in several sectors owing to their well-established capacity to process visual input. This study presents a methodology for classifying and identifying objects using the YOLOv4 object detection algorithm. Convolutional neural networks (CNNs) have shown exceptional efficacy in the tasks of object tracking and feature extraction from pictures. Therefore, the enhanced network architecture optimizes both the precision of identification and the speed at which it operates. This research will contribute to developing mixed-reality simulations system for object detection and tracking in collaborative environment that are accessible to everyone, including users in the architectural filed. The model was evaluated in comparison to other object detection approaches. Based on the empirical results, it was observed that the YOLOv4 model exhibited a mean average precision (mAP) of 0.988, surpassing the performance of both YOLOv3 and other object identification models.

Keywords: Mixed reality, YOLOv4, Artificial intelligence, object detection.

1. Introduction

During the latter part of the 90s, Mixed Reality (MR) gained significant traction among the scientific community, with its inherent capabilities and practical uses in the physical realm being well acknowledged. The concept of augmenting our comprehension of the world by the act of observation has fostered scholarly exploration in this domain, hence proposing several pragmatic implications. MR has emerged as a prominent and rapidly growing immersive experience in the 21st century. MR has significantly transformed several sectors, such as healthcare, education, tourism, design, manufacturing, and related businesses. The widespread use of MR has led to its remarkable development at an unparalleled pace [1][2][3]. In today's rapidly evolving landscape of combat training and military preparedness, the integration of cutting-edge technologies has become a paramount requirement. One such groundbreaking innovation is the development of a Mixed Reality (MR) and Artificial Intelligence (AI) based remote collaboration platform designed to revolutionize real-time combat training [4]. This transformative platform brings together the realms of physical and digital worlds, offering an unparalleled training experience for military personnel. By combining the immersive capabilities of mixed reality with the cognitive power of AI, this platform redefines how combat training is conducted, fostering enhanced situational awareness, decision-making skills, and teamwork, all within a secure and adaptable environment [5][6]. MR-based applications are one of the top 10 ranked ICT technologies in 2020 [7]. Numerous investigations have been undertaken to examine the field of MR technology, resulting in the development of several survey categories pertaining to this technological domain [8][9].

In recent study, video and teleconferencing technologies are already in widespread use, which enables researchers to concentrate their efforts on developing more innovative alternatives that make use of VR, and MR technology [10]. The infrastructure necessary for remote collaboration in the realms of virtual reality, and MR is slowly but gradually being developed, but it is not yet fully developed [11][12]. The construction of such a system requires expertise in a wide variety of different fields, including as 3D modelling, animation, the creation of interactive systems, the development of avatars and dynamic content, multi-user interactions, and a great deal more. This is because such a system is comprised of a large number of distinct components and features. In addition, mixed reality systems are often implemented by employing a mixture of MR gear and VR technology [13]. This demands a certain level of competence in both of these technologies. It is vital to collaborate in a variety of professional fields, such as healthcare and mechanical maintenance, as well as professional education. People needed to be in the same physical location in order to carry out the majority of the tasks that required collaboration in the past. On the other hand, the future of cooperation is becoming more digital and scattered across time and location. For cooperation to be durable, scalable, and successful, the capacity to collaborate from a distance is necessary [14]. There are several challenges that need to be addressed before a mixed reality and AI based remote collaboration platform for real-time combat training can be realized. One challenge is the need for high-bandwidth networks to support the real-time transmission of data. Another challenge is the need for lightweight and affordable MR headsets that are suitable for military use. Despite these challenges, the potential benefits of this type of platform are significant. It could help to improve the training of soldiers, reduce the cost of training, and make training more efficient and realistic.

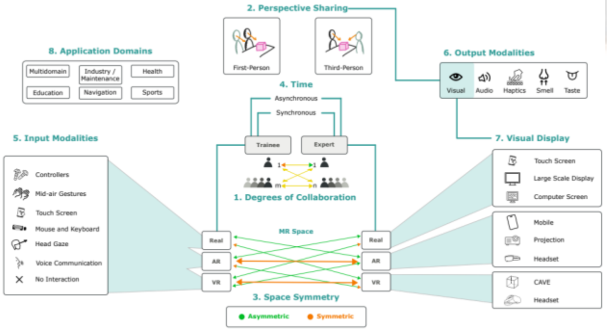


Figure 1. Taxonomy for remote assistance and training in MR environments [15].

The remaining of this paper structured as follows. Section 2 confers the details reviewed of virtual reality, mixed reality with tracking in collaborative environment using AI. Section 3 described the problem formulation, and section 4 explained about methodology that how our methods is effective to enhance the object detection in collaborative environment. Section 5 highlight the results of the study and dataset while section 6 depicts the conclusion and future scope of our study.

2. Review of Literature

Zechner et al., (2023)[16] examined that VR has significant prospects for police personnel to engage in DMA training inside cognitively challenging and high-pressure scenarios. And provided an overview of the findings derived from a three-year study, including the gathering of needs from seasoned police trainers and industry experts, as well as the presentation of quantitative and qualitative outcomes from human factor studies and field testing. The results of this study reveal several benefits associated with VR training. One advantage is the ability to safely simulate high-risk scenarios within controlled and repeatable training environments. VR training also offers the opportunity to incorporate a diverse range of avatars that would be impractical to utilize in real-life training, such as individuals from vulnerable populations or animals. Additionally, VR training enables the handling of hazardous equipment, such as explosives, without actual risk. However, it is important to acknowledge the challenges associated with VR training, including issues related to tracking, locomotion, and the development of intelligent virtual agents. Also emphasized the need of establishing a robust correspondence between training didactics and technical capabilities. Additionally, it presents alternative methods to address this issue. Moreover, the training results exhibit transferability to practical police responsibilities and may be applicable to other fields that might get advantages from simulation-based training.

Alatawi et al., (2023)[17] stated that the usage of MR technology is on the rise in the maintenance sector because it has the potential to increase productivity while simultaneously lowering expenses. This is accomplished via the provision of real-time direction and instruction to employees while they do repairs and maintenance chores. Image-based or object-based tracking must be calibrated in order to apply mixed reality for work aids in mobile apps. This study introduces a novel approach that utilizes MR and deep reinforcement learning (RL) techniques to develop a model for training and maintaining NanoDrop Spectrophotometers. This system may be utilized for speedy repair processes in an Industry 4.0 (I4.0) scenario. The system employs a camera to do feature matching, tracking algorithms, and 3D modelling in order to locate the item that is the focus of its attention. Once the detection has been finished, the device used by the maintenance operator will get instructions that are crystal clear and simple to comprehend thanks to mixed reality technology. The outcomes of the study indicate that the model's goal strategy produced a reward with a mean value of 1.0001.000 and a standard deviation value of 0.0000.000. This indicates that all of the benefits that might be received via the activity or setting in question were precisely the same. The fact that there is no fluctuation in the results reveals that the reward standard deviation is 0.0000.000, which is the lowest possible value.

Lee et al., (2022)[18] studied as a result of the COVID-19 epidemic, there has been a transition from traditional face-to-face instruction to remote education, wherein the majority of students are now engaging in virtual learning using video conferencing platforms. This alteration hinders the engagement of pupils in active involvement during class. Video-based education has inherent limits when it comes to substituting practical courses, since the latter need the acquisition of both theoretical understanding and hands-on experience. Also present a proposed system that integrates virtual reality and metaverse techniques inside the educational setting, aiming to address the limitations of current distant practical education models. In accordance with the suggested framework and developed an Aeroplan maintenance simulation and then executed an empirical investigation to compare the efficacy of our system against a video-based teaching approach. In order to assess the efficacy of education, examinations were administered to evaluate knowledge acquisition and retention, while the extent of engagement was examined via the analysis of survey data. The experimental findings indicate that the cohort using the suggested approach exhibited superior performance compared to the video training group in terms of their scores on knowledge assessments. Consequently, the appropriateness of the suggested system's usability was evaluated.

Estrada et al., (2022)[19] focused in the field of object identification, deep learning (DL) methods have shown to perform very well. Meanwhile, mixed reality (MR) methods are revolutionizing our professional and interpersonal interactions. Devices like the multimeter, oscilloscope, wave generator, and power supply are all in the realm of the mixed reality program's automatic object recognition feature, which is driven by deep learning. Using TensorFlow's object identification API, deploy a deep neural network model for machinery detection called MobileNet-SSD v2. The screen will show the relevant mixed reality instruction when a piece of equipment is recognized. The developed equipment detection model has an average recall of 85.3% and an average mean average precision (mAP) of 81.4%. Also provide a multimeter lesson in which digital models are overlaid on physical multimeters to show how the suggested framework might be put to use in the real world. The primary objective of this course is to use the Unity3D game engine as the primary tool for combining DL (Deep Learning) and MR frameworks, hence facilitating the creation of virtual worlds. The proposed framework exhibits the capacity to function as the fundamental structure for MR and machine learning (ML) frameworks, with potential applications in both business and academic education settings.

Mao et al., (2021)[20] suggested that the field of MR has come a long way in the last decade, finding use in many different fields like education, industry, medicine, and even the military. There has been an uptick in investment towards developing better battlefield observation systems and simulating training environments for troops. Tactical training using mixed reality is a new area of study with enormous promise. Research in this area has shown both the potential and the limits of MR technology for use in the military. In order to evaluate and compare the results of the experiment, an independent sample "t" was put through a series of tests. The testing findings of the MR-based tactical education system demonstrate a considerable shift in how students conceptualize combat scenarios and tactical operations. The prototype system and experimental material contribute significantly to tactical instruction at the Army Academy and serve as a helpful benchmark for future research and development of mixed reality (MR) for military usage.

Binsch et al., (2021)[21] presented research investigates whether or not varying sources of stress during a VR military training scenario lead to elevated levels of physiological stress. This would provide credence to the practice of using VR modelling for stress training and the educational use of physiological monitoring of students. To test their knowledge of their surroundings, 63 cadets participated in a patrol scenario. There were four stages in which stressors were progressively introduced. The cadets' ECGs, BPs, EDAs, cortisol levels, and subjective assessments of threats and challenges were also recorded. Using metrics like heart rate, HRV, and EDA, we determined that only the first phase significantly elevated physiological stress. Even though the third and fourth phases were intended to be the most demanding, physiological stress seemed to remain high throughout each of them. Subjective threat/challenge evaluation scores were considerably higher for the cadets identified as threat responders (n = 3) compared to the cadets classified as challenge responders (n = 21). Other imposed stressors did not generate a significant physiological reaction in the assessed VR training scenario, suggesting that the novelty of the scenario was the sole effective stress trigger. And conclude that virtual reality (VR) training scenarios designed to hone a person's ability to handle pressure should put them in situations where they must adapt to unforeseen but relevant challenges.

Malik et al., (2020)[22] intended that there has been a lot of focus on human-machine interaction as a path towards hybrid automation in production. The complementary relationship between humans and robots is fostering a new kind of hybrid automation. Designing and re-designing human-robot collaboration (HRC) systems for use in assembly is difficult because of the high standards of flexibility, adaptability, and safety that must be met. In order to facilitate the design of complicated HRC systems, time-based continuous simulations may provide a risk-free virtual area for testing and validation. However, traditional simulations don't provide a way to immerse oneself in the role of an end-user of the future manufacturing system. This study investigates the evolution of VR technology for human-centered production system design and creates a unified framework for incorporating VR into human-robot simulation. As an event-driven simulation, it was useful for calculating human and robot cycle durations, creating a process plan, optimizing the layout, and writing a programme to operate the robot. The same modelling is utilized in virtual reality to control manufacturing machinery, most notably the robot. In addition, a virtual robot is developed in the AWS Sumerian environment to aid the VR designer.

Park et al., (2020)[23] studied that smart glasses that use mixed reality (MR) technology have found use in settings as diverse as classroom instruction, facility upkeep, and teamwork. Yet, prior research on wearable mixed reality (MR) technology has been constrained in its capacity to provide efficient support in situational activities, mostly due to its dependence on static visualization and registration based on MR markers. This research paper introduces a new approach to providing support for tasks by integrating deep learning methods for object identification and instance segmentation with Mixed reality (MR) technology, enabling users to get hands-free and effective visual guidance. In this particular circumstance, the Mask R-CNN is used for the objective of instance segmentation. In addition, the technique of marker less mixed reality is used in combination with the Mask R-CNN algorithm to superimpose a three-dimensional spatial representation of a genuine item onto its physical surroundings. To improve the user's capacity to recognize and comprehend tangible entities while navigating in the physical surroundings, the integration of 3D spatial data with instance segmentation is used to provide assistance and navigation for activities involving three-dimensional items. The 3D annotation and cooperation between multiple employees is also supported by 2.5D or 3D copies, even without prepared 3D models. That way, users can simulate dynamic production settings more accurately.

Boletsis et al., (2019)[24] studied the recent breakthroughs in technology and interaction within the VR domain have ushered in a new age, not only for VR itself, but also for VR locomotion. In the contemporary period, widely recognized and widely used VR locomotion approaches serve as primary reference points for evaluating and assessing the performance of novel VR locomotion systems. Simultaneously, there exists a need for more exploratory and comparative investigations into modern VR locomotion strategies. These studies aim to chronicle the distinctive characteristics of interaction associated with these techniques and provide guidance for the development of novel approaches. This paper provides an empirical assessment research that compares modern and widespread VR locomotion systems, with a focus on analyzing the user experience (UX) they provide. Initially, the predominant VR locomotion approaches are discerned by an examination of relevant scholarly literature. These techniques include walking-in-place, controller/joystick, and teleportation. The research consists of a cohort of twenty-six adult participants who engage in a task resembling a game, using various strategies. The results imply that walking in place offers the most immersion, although also causing the most psychophysical suffering. The ease of use of controller/joystick VR locomotion may be attributed to the users' experience with such input devices. On the other hand, teleportation is often regarded as an excellent method of navigating in virtual reality owing to its ability to swiftly transport users to other locations. However, it is worth noting that teleportation might disrupt the users' feeling of immersion due to the visual discontinuity caused by the abrupt transitions.

2.1 Comparison of reviewed technique

There is a wide range of author who studied on the mixed reality and AI based remote collaboration platform for real time combat training and give their findings as shown in table 1.

Table 1. Comparison of reviewed technique

Authors [Ref.]	Technique	Outcome
Zechner et al., (2023) [16]	DMA and VR	The training results exhibit transferability to practical police responsibilities and may be applicable to other fields that might get advantages from simulation-based training.
Alatawi et al., (2023) [17]	MR and RL	The outcomes of the study indicate that the model's goal strategy produced a reward with a mean value of 1.0001.000 and a standard deviation value of 0.0000.000.
Lee et al., (2022) [18]	Virtual learning	The experimental findings indicate that the cohort using the suggested approach exhibited superior performance compared to the video training group in

		terms of their scores on knowledge assessments.
Estrada et al., (2022) [19]	DL and MR	The suggested framework has the potential to serve as the backbone of MR and ML-based frameworks for application in corporate and academic education.
Mao et al., (2021) [20]	3D MR technology	The testing findings of the MR-based tactical education system demonstrate a considerable shift in how students conceptualize combat scenarios and tactical operations.
Binsch et al., (2021) [21]	VR	And conclude that virtual reality (VR) training scenarios designed to hone a person's ability to handle pressure should put them in situations where they must adapt to unforeseen but relevant challenges.
Malik et al., (2020) [22]	HRC system	As an event-driven simulation, it was useful for calculating human and robot cycle durations, creating a process plan, optimizing the layout, and writing a programme to operate the robot.
Park et al., (2020) [23]	CSCW	By means of discerning emerging patterns, we provide prospective avenues for further study in the field of Mixed Reality (MR).
Boletsis et al., (2019) [24]	VR	The results imply that walking in place offers the most immersion, although also causing the most psychophysical suffering.

3. Problem formulation

In the contemporary world, the convergence of Artificial Intelligence (AI) and Mixed Reality (MR) technologies has the potential to revolutionize the way we interact with our physical and digital surroundings. However, despite significant advancements in both AI and MR, there exists a pressing need to address the challenge of real-time object detection and tracking within collaborative MR environments. This problem statement outlines the core issues and motivations for research in this domain:

Limited Real-Time Object Detection: Current AI algorithms have made substantial progress in object detection, but there are still challenges in achieving real-time performance, especially in complex MR environments. Efficient real-time detection is crucial for enabling seamless interactions and experiences in these collaborative spaces.

Collaborative Environments: The rising demand for collaborative MR environments in fields such as education, healthcare, engineering, and entertainment requires robust solutions for shared object detection and tracking. Collaborative scenarios involve multiple users sharing a physical space and their digital counterparts, making accurate tracking essential for user engagement and effective communication.

Mixed Reality Challenges: The dynamic nature of mixed reality, where the digital and physical worlds merge, presents unique challenges for object detection and tracking. These challenges include occlusions, changes in lighting, unpredictable movements, and the need for precise spatial alignment between digital and physical objects.

Safety and Accuracy: In applications where safety is paramount, such as medical procedures or industrial settings, accurate object detection and tracking are critical to prevent accidents and ensure that users can rely on the digital information overlaying the real world.

Cross-Domain Applications: Object detection and tracking in collaborative MR environments have applications across various domains, from gaming and entertainment to training and education, remote assistance, and telemedicine. Thus, addressing this problem has the potential to benefit a wide range of industries and users.

In light of these challenges and opportunities, this research paper aims to investigate and propose innovative solutions that combine AI and MR technologies to enable real-time object detection and tracking in collaborative mixed reality environments.

4. Research methodology

Earlier research conducted in the realm of remote collaboration were constrained by either the use of two-dimensional depictions of the immediate work environment or the confinement of a restricted workspace. In the course of our investigation, we attempted to address these deficiencies within a single system. Artificial intelligence (AI) plays a crucial role in mixed reality (MR) devices by enhancing various aspects of the MR experience, including object detection and interaction. Here's how AI works with MR devices and its connection to object detection:

Sensor Data Processing: MR devices are equipped with various sensors such as cameras, depth sensors, and microphones. These sensors capture real-world information, including the environment, objects, and people. AI algorithms process the sensor data to extract relevant features and information.

Computer Vision: AI-powered computer vision algorithms analyze the sensor data to identify objects, surfaces, and people in the physical world. Computer vision techniques can detect the shape, size, position, and movement of objects in the MR environment.

Object Detection: Object detection is a specific computer vision task that involves identifying and locating objects within images or video frames. AI models, such as convolutional neural networks (CNNs), are used to perform object detection in real time on MR devices. These models can recognize objects, track their positions, and even estimate their 3D spatial coordinates.

Tracking and Augmentation: AI algorithms not only detect objects but also track their movements and positions as users interact with them or move around. This tracking information is essential for seamlessly integrating virtual objects into the MR environment and ensuring they remain in the correct physical locations.

Integration of Virtual Objects: AI combines object detection with spatial understanding to seamlessly integrate virtual objects into the real world. Virtual objects can interact with the physical environment, cast shadows, and respond to lighting conditions. The real-time tracking of physical objects and the user's perspective allows virtual objects to stay in the user's view and maintain a consistent relationship with their surroundings.

Contextual Interactions: AI uses object detection to understand the context of the MR environment. For example, it can identify a table and place a virtual object on top of it or recognize a user's hand to trigger a virtual control interface. The context-aware AI can also provide information about recognized objects, enhancing the user's understanding of the physical environment.

Safety and Collision Avoidance: Object detection, combined with AI, can help prevent collisions or unsafe interactions in the MR environment. For instance, AI can alert the user or adjust virtual object positions to avoid real-world obstacles.

The connection between AI and object detection in MR devices is essential for creating immersive and interactive mixed reality experience. The combination of computer vision, sensor data processing, and AI algorithms allows MR devices to understand the physical environment, identify objects, and create interactive and immersive mixed reality experiences. This technology is used in various applications, including augmented reality (AR) glasses, virtual reality (VR) headsets, and MR software platforms for gaming, education, healthcare, and industrial use cases.

The remaining procedure of the work is described given below.

4.1 Virtual environment

The efficacy of the suggested methodology is often influenced by factors such as the backdrop, direction, and lighting conditions. The suggested approach for generating synthetic pictures incorporates the use of the following characteristics:

The selection of the number of unique MR markers seen in each scene is determined randomly from the set of teachable categories.

- The likelihood of seeing one or more MR markers in the same scene is set at 50%.
- The scaling range of the MR marker is randomly determined to be between 20% and 40% of the total area of the scene.
- The rotation angle of the MR marker is randomly assigned within the range of 0 to 360 degrees with respect to the z-axis of the scene.
- A diverse range of lighting sources is produced by the use of random camera perspectives in order to showcase various real-world lighting effects.

These variants include a diverse array of real-world circumstances and external factors that have the potential to enhance the accuracy of marker recognition.

The applicability of object detection and tracking using mixed reality, especially with Microsoft HoloLens, is significant and has practical implications across various domains. Here's a summary of their applications: HoloLens, with its object detection capabilities, can assist technicians in identifying and troubleshooting machinery by overlaying relevant information on physical equipment in real-time. Object detection and tracking can be used to create realistic training simulations where users interact with virtual representations of objects, enhancing the learning experience for tasks such as equipment assembly or repair.

4.2 Microsoft lens

This section presents the chosen Head-Mounted Display (HMD) and provides a concise description of the algorithms used by the Mixed Reality (MR) system described given below.

The Microsoft HoloLens, often known as HoloLens, is a head-mounted display (HMD) that has the capability to project holograms into a physical environment. The device is equipped with many types of sensors, including RGB and depth cameras, as well as an Inertial Measurement Unit (IMU). These sensors enable the device to perceive neighboring forms and objects in real time. The device is equipped with two computing units integrated on-board, namely an Intel 32-bit CPU and a Holographic Processing Unit (HPU). These units are

***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

responsible for managing both the operating system and the real-time updating of holograms. The aforementioned technology has the capability to ascertain its relative location in relation to the physical world, while concurrently generating a real-time map of the immediate surroundings. By using its inherent capabilities, those who use this headgear have the ability to navigate unrestricted across physical space, hence enabling the collection and analysis of many data types that may be immediately visualized via the glasses. Nevertheless, the computing capacity of the system is constrained, which might potentially hinder its ability to perform computationally intensive algorithms in real-time. In order to achieve this objective, it is advisable to take into consideration the use of an external server equipped with GPU processing capabilities.

4.3 Object Tracking

When taking into account a changing environment, the process of object detection is only the first phase of the whole work. There is a pressing need for a rapid approach that can effectively track the dynamic positions of several objects in real-time. The tracking methods calculates the updated location of the item through contrasting the most recent frame with the previous one. It can be inferred that latency is not permissible in this activity, therefore necessitating the selection of an algorithm with minimal complexity that can be implemented immediately by HoloLens. The use of a suitable tracking approach restricts the need to identify a novel entity only to instances when an unfamiliar item emerges inside the Field of View (FoV), hence enhancing the overall efficiency and speed of the system. Position tracking methods for wireless sensor networks/IoT is also suggested in [25][26].

4.4 Object detection techniques

The categorization of object identification using a CNN may be divided into two distinct types: area nomination and regression. Region nomination algorithms, such as R-CNN and Fast R-CNN, use a step-by-step detection method algorithm. The first step involves the extraction of proposal areas from the pictures using a method known as selective search. Subsequently, the images are classified within these proposal regions. However, there is a significant reduction in the frame per second rate. The YOLO algorithm uses the regression technique to make predictions about the object's bounding box and class label, as opposed to using the suggested region approach. Nevertheless, the detection accuracy experiences a decline as the frame rate rises, owing to the simpler network topology.

- **YOLOv4**

YOLOv4 is the fourth iteration within the lineage of models known as You Only Live Once. The priority of real-time detection is emphasized, and training is performed on a single Graphics Processing Unit (GPU). The input consists of a picture and the corresponding numerical values representing the bounding boxes. The input picture is partitioned into a grid composed of square cells, with each item being recognized inside the grid cell that encompasses its center For every grid cell, predictions will be made for the B bounding boxes along with their respective confidence ratings. The rating provided indicates the level of accuracy of the box and the level of certainty of the model on the presence of the object inside the box. The confidence number is determined by calculating the Intersection over Union (IoU) between the anticipated values and the actual values of the boxes. The study conducted by YOLO [27][28] reframes the problem as a regression task and effectively forecasts the bounding box attributes of the image as well as the object's state. The structure consists of several divisions, as seen in Figure 2.

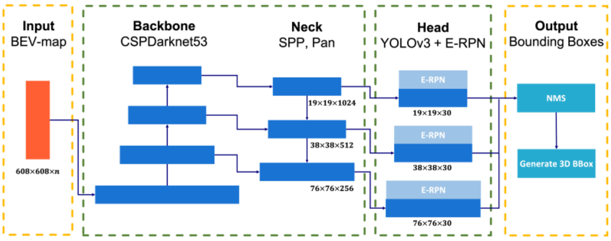


Figure 2. The architecture of YOLOv4

The training pictures in the input set are subsequently processed by the network, and then trained in parallel batches on the GPU. The grouping and extraction of characteristics are carried out by the Backbone (CSPDarknet53) and the Neck (SPP and PAN) components. The combined components of the Neck and Head constitute the Object Detector, with the Head responsible for the final detection and prediction process. The individual in the position of Head is accountable for the identification and categorization of objects or phenomena, including both the determination of their specific location and their classification. The YOLO algorithm utilizes a grid structure, specifically a S x S grid, to partition the picture into cells. Each cell inside this grid serves as a container for each item's detection point, which is represented by the center of the object. The prediction array for each grid cell has five items, namely the center coordinates of the bounding box, its width and height dimensions (w and h), and the confidence level associated with it, as seen in Figure 2. As seen in Figure 2, the Yolo approach partitions the whole picture into cells with fixed dimensions of 19 by 19. Subsequently, each individual cell will have the responsibility of predicting its own bounding rectangle, so providing a heightened degree of accuracy. Non-maximum suppression is a technique used to merge bounding rectangles by evaluating their overlapping area and predicting the bounding rectangle that best represents a recognizable picture.

***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

4.5 Proposed methodology

Our proposed system uses only images acquired from a monocular camera as input and does not use a model-independent application programming interface such as ARKit or ARCore. Therefore, it contributes to the realizations of a simulation system that is easy for everyone to use, including users in the architectural field. The execution process of the tracking system proposed in this study can be divided into preprocessing and main processing. Preprocessing defines the tracking reference points and performs the initial camera alignment. Figure 3 depict a schematic diagram and flowchart of the main process in our proposed system.

Step 1: The first step in the main process is to acquire input from the webcam. Next object detection is performed on the input image and the results are recorded.

Step 2: Then, we focus on the tracking references points inside the bounding boxes of the objects detected in the image of the previous and current frame. These are excluded once from the processing that is to be performed later, and the information of their 3D coordinates is saved. By comparing all of the tracking references points (2D coordinates on the mixed reality execution screen) to the bounding boxes of every detected object (maximum and minimum 2D coordinates). The system determines whether the tracking references point is inside the bounding box.

Step 3: In this phase, optical flow estimation is performed for the tracking references points outside of the bounding box, and their positions on the mixed reality (MR) execution screen are tracked frame by frame.

Step 4: In this step, the current camera pose is then estimated from the correspondence between the 2D and 3D coordinates of the tracking reference points that remain after excluding those that contained large errors in the optical flow estimations.

Step 5: Finally, from the estimated camera pose and the saved 3D coordinates of the tracking references point inside the bounding boxes, their 2D coordinates in the current frame are obtained and returned as the target for the tracking process in the next frame. This process makes it possible to secure the tracking reference point while maintaining the robustness of the tracking system, as it is less susceptible to obtain errors from moving objects.

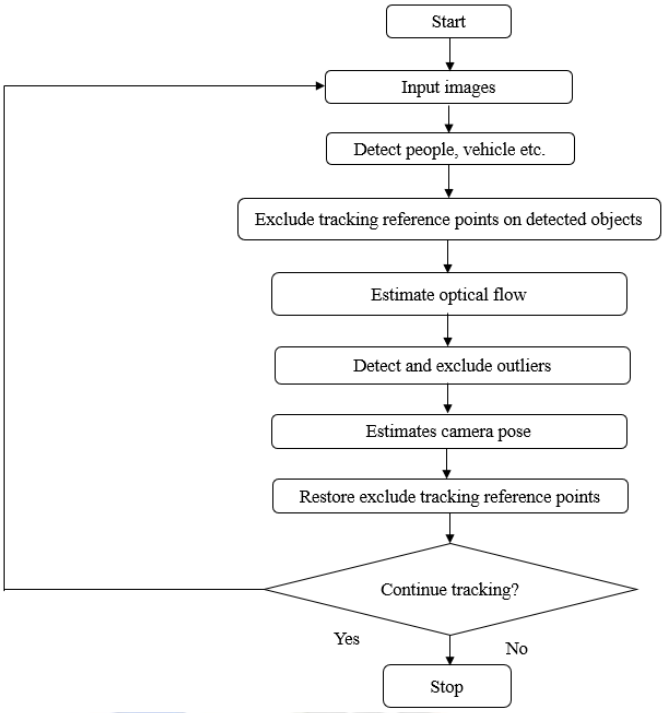

Start
Input images
Detect people, vehicle etc.
Exclude tracking reference points on detected objects
Estimate optical flow
Detect and exclude outliers
Estimates camera pose
Restore exclude tracking reference points
Continue tracking?
Yes
No
Stop

Figure 3. Flowchart of the proposed tracking method.

5. Dataset

The first evaluation of YOLOv4 was conducted using the Microsoft COCO dataset, which is well recognized for its comprehensive representation of common items in various contexts. For examples, the COCO dataset is an extensive collection of data used for object recognition, segmentation, and captioning tasks [29]. The MS-COCO dataset is well recognized and used for evaluating the efficacy of various object identification techniques. The COCO dataset is often used for the purpose of training object detection algorithms. The dataset includes bounding box coordinates for a diverse range of 80 item categories. These coordinates may be used to train models for the purpose of detecting bounding boxes and accurately classifying objects inside pictures. The COCO dataset has annotations for instance segmentation, making it suitable for training models in this specific purpose. The COCO dataset is well-suited for semantic segmentation tasks because to its extensive collection of images accompanied with pixel-level annotations for each class included within the images.

MS COCO offers various types of annotations,

- **Object detection** with bounding box coordinates and full segmentation masks for 80 different objects.
- **Stuff image segmentation** with pixel maps displaying 91 amorphous background areas.
- **Panoptic segmentation** identifies items in images based on 80 "things" and 91 "stuff" categories.

***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

- **Dense pose** with over 39,000 photos featuring over 56,000 tagged persons with a mapping between pixels and a template 3D model and natural language descriptions for each image.
- **Key Point** annotations for over 250,000 persons annotated with key points such as the right eye, nose, and left hip.

6. Result and discussion

Computer science and artificial intelligence (AI) professionals often use the COCO dataset in relation to various research challenges. The investigation of this Object Detection Model involves the use of the MS COCO data, which is a well-recognized dataset for object identification and segmentation, in conjunction with a labelling dataset recommended by Microsoft. The COCO picture collection, selected for its emphasis on enhancing image recognition, is representative of Common Objects in our area. The results of this study are described given below.

6.1 Performance metrices

Mean average precision: The mean average accuracy measure is often used in the evaluation of deep neural networks that are based on computer vision techniques. The average precision metric calculates the mean accuracy value throughout the range of recall values from 0 to 1.

$$\text{Precision} = TP/TP+FP$$

TP represents genuine positive, which refers to the accurate identification of positive predictions. The acronym FP denotes false positives, which refers to instances when positive forecasts are inaccurate.

Recall: Total up the number of true positive and accurate hits that were found in the search results. Recall is a high degree of accuracy may be defined as a high percentage of correct hits being returned, or the number of positives. The formula is shown in the following as:

$$\text{Recall} = TP/TP+FN$$

Detection Speed: Another crucial assessment parameter for real-time object detection is the detection accuracy.

Frames Per Second (FPS): The frames per second (FPS) is a significant metric for quantifying the speed of detection, since it represents the number of objects that the algorithm can see within one second. Yolov4 provided us with the necessary levels of average accuracy and FPS.

Table 1 shows the performance of tracking using the YOLO models.

Table 2. Performance of tracking using YOLO models

Models	FPS	Detection speed
YOLOv2	40.6	173.54ms
YOLOv3 608	60.2	19.67ms
YOLOv3 416	61.3	18.23ms
YOLOv4	110.56	16.01ms

Various experiments were undertaken in order to assess the suggested approach. Figure 4 displays the outcomes of the training loss, whereas Figure 5 presents the mean average precision values, namely mAP and mAP50. The findings demonstrate that the YOLOv4 model achieves a mAP of over 80% and a mAP of over 55% when applied to our dataset.

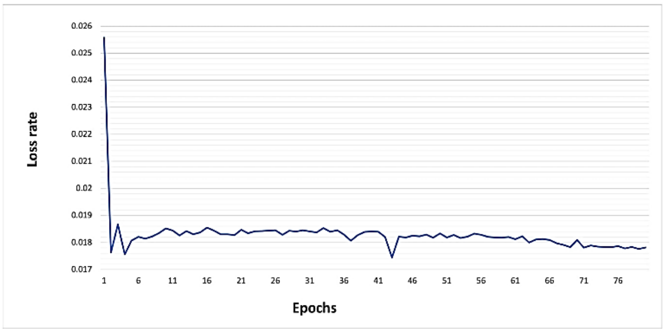


Figure 4. The loss curve during training of YOLOv4.

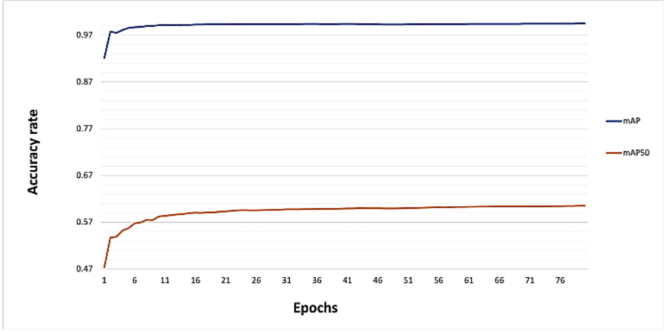
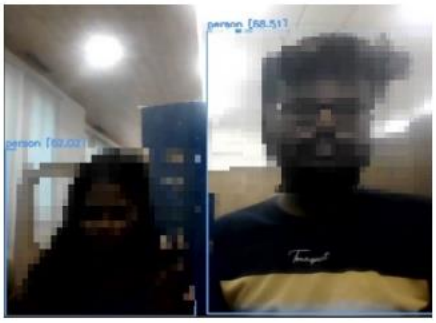


Figure 5. YOLOv4 training (mAP) and (mAP50) curves.

The analysis of the captured images obtained from our camera was documented in Figure 6 (a) and Fig6. (b). Additionally, the corresponding classes and their respective confidence values were assigned.

(a)



(b)

Figure 6. (a) Outcome of images using YOLOv4, (b) object ensnared within human images

6.2 Comparative analysis

Additionally, we conducted a comparative analysis of our dataset using the same YOLO family model as shown in table 3. The dataset exhibits enhanced performance metrics while preserving a detection time of 0.01s, aligning with the established real-time detection criterion. Remarkably, the mean Average Precision (mAP) and mAP at 50% intersection over union (mAP50) values saw an approximate twofold increase upon using our dataset for training deep neural networks. The findings of this study suggest that the strategy described in this research has efficacy in mitigating the presence of outliers and noise, which have the potential to adversely impact the accuracy of predictions. Table 3 depict the comparison of the accuracy results using YOLO models as well as in Figure 7. From table 3, YOLOv2 model attained 72.1% precision, recall is 75.42% and mAP is 44.0% and YOLOv3 608 is slightly increased precision, recall, mAP and mAP50 as compared to YOLOv2. Therefore, YOLOv4 models is attained high precision, recall, mAP, and mAP50 as compared to other models.

Table 3. Comparison of the accuracy results using YOLO models.

Models	Precision	Recall	mAP	mAP50
YOLOv2	72.1%	75.42%	44.0%	21.6%
YOLOv3 608	59.3%	50.6%	57.9%	33.0%
YOLOv3 416	60.5%	58.9%	55.3%	31.0%
YOLOv4	98.88%	99.31%	99.5%	95.6%

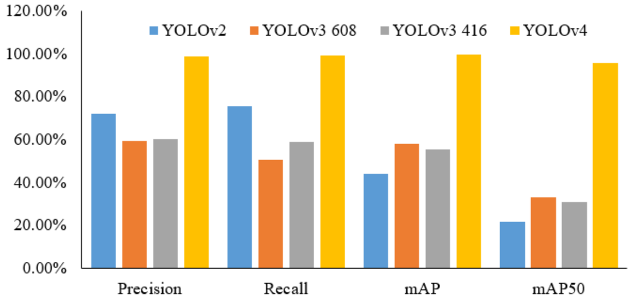


Figure 7. Comparison graph with existing methods

7. Conclusion and future work

The use of MR has garnered considerable interest due to its potential efficacy as a visual tool in landscape research. It is essential for MR (Mixed Reality) technology to accurately register both the physical and virtual environments in order to provide an immersive experience to the user. One of the primary contributors to location inaccuracies arises from the monitoring of the camera's real-time stance inside a physical environment. The application prototype was created with ARKit as the major software framework, including YOLOv4 as the principal object detection model within the DNN component. In this study, the authors propose a unique framework named YOLOv4 for the purpose of real-time object recognition. This framework employs optimized algorithms for the loss function and object augmentation, building upon the foundation of YOLOv4. The efficacy of YOLOv4 is evaluated in relation to existing models. The results indicate that our model effectively addresses the challenge presented by a complex backdrop, and the speed at which it detects objects is satisfactory for meeting the real-time requirements of the classifier. YOLOv4 is a compacted iteration of several computer vision techniques used for the purpose of object detection. However, it is possible to expand the suggested technique to specifically target the detection of items in close proximity, since each grid has the potential to suggest two bounding boxes.

References

[1]. Berciu, Alexandru G., Eva H. Dulf, and Iulia A. Stefan. "Flexible Augmented Reality-Based Health Solution for Medication Weight Establishment." Processes 10, no. 2 (2022): 219.

[2]. Sırakaya, Mustafa, and Didem Alsancak Sırakaya. "Augmented reality in STEM education: A systematic review." Interactive Learning Environments 30, no. 8 (2022): 1556-1569.

[3]. Chouchene, Amal, Adriana Ventura Carvalho, Fernando Charrua-Santos, and Walid Barhoumi. "Augmented reality-based framework supporting visual inspection for automotive industry." Applied System Innovation 5, no. 3 (2022): 48.

[4]. Flavián, Carlos, Sergio Ibáñez-Sánchez, and Carlos Orús. "The impact of virtual, augmented and mixed reality technologies on

the customer experience." Journal of business research 100 (2019): 547-560

[5]. Roh, Byeong-hee, Geunkyung Choi, Seungwoon Lee, S. J. Kim, and Jinsuk Kang. "Mixed Reality-Enabled Multilateral Collaboration Application Platform with AI and IoT Convergence." In 2023 IEEE International Conference on Consumer Electronics (ICCE), pp. 1-4. IEEE, 2023.

[6]. Rizzo, Albert 'Skip, and Russell Shilling. "Clinical virtual reality tools to advance the prevention, assessment, and treatment of PTSD." European journal of psychotraumatology 8, no. sup5 (2017): 1414560.

[7]. Makhataeva, Zhanat, and Huseyin Atakan Varol. "Augmented reality for robotics: A review." Robotics 9, no. 2 (2020): 21.

[8]. Lange, Belinda, Sebastian Koenig, Chien-Yen Chang, Eric McConnell, Evan Suma, Mark Bolas, and Albert Rizzo. "Designing informed game-based rehabilitation tasks leveraging advances in virtual reality." Disability and rehabilitation 34, no. 22 (2012): 1863-1870.

[9]. Zahabi, Maryam, and Ashiq Mohammed Abdul Razak. "Adaptive virtual reality-based training: a systematic literature review and framework." Virtual Reality 24 (2020): 725-752.

[10]. Robitaille, Nicolas, Philip L. Jackson, Luc J. Hébert, Catherine Mercier, Laurent J. Bouyer, Shirley Fecteau, Carol L. Richards, and Bradford J. McFadyen. "A Virtual Reality avatar interaction (VRai) platform to assess residual executive dysfunction in active military personnel with previous mild traumatic brain injury: proof of concept." Disability and Rehabilitation: Assistive Technology 12, no. 7 (2017): 758-764.

[11]. Van Krevelen, D. W. F., and Ronald Poelman. "A survey of augmented reality technologies, applications and limitations." International journal of virtual reality 9, no. 2 (2010): 1-20.

[12]. Scavarelli, Anthony, Ali Arya, and Robert J. Teather. "Virtual reality and augmented reality in social learning spaces: a literature review." Virtual Reality 25 (2021): 257-277.

[13]. Vaughan, Neil, Bodgan Gabrys, and Venketesh N. Dubey. "An overview of self-adaptive technologies within virtual reality training." Computer Science Review 22 (2016): 65-87.

[14]. Marion, Tucker J., and Sebastian K. Fixson. "The transformation of the innovation process: How digital tools are changing work, collaboration, and organizations in new product development." Journal of Product Innovation Management 38, no. 1 (2021): 192-215.

[15]. Fidalgo, Catarina G., Yukang Yan, Hyunsung Cho, Maurício Sousa, David Lindlbauer, and Joaquim Jorge. "A Survey on Remote Assistance and Training in Mixed Reality Environments." IEEE Transactions on Visualization and Computer Graphics 29, no. 5 (2023): 2291-2303.

[16]. Zechner, Olivia, Lisanne Kleygrewe, Emma Jaspaert, Helmut Schrom-Feiertag, RI Vana Hutter, and Manfred Tscheligi. "Enhancing Operational Police Training in High Stress Situations with Virtual Reality: Experiences, Tools and Guidelines." Multimodal Technologies and Interaction 7, no. 2 (2023): 14.

[17]. Alatawi, Hibah, Nouf Albalawi, Ghadah Shahata, Khulud Aljohani, A'aeshah Alhakamy, and Mihran Tuceryan. "Augmented Reality-Assisted Deep Reinforcement Learning-Based Model towards Industrial Training and Maintenance for NanoDrop Spectrophotometer." Sensors 23, no. 13 (2023): 6024.

[18]. Lee, Hyeonju, Donghyun Woo, and Sunjin Yu. "Virtual reality metaverse system supplementing remote education methods: Based on aircraft maintenance simulation." Applied Sciences 12, no. 5 (2022): 2667.

[19]. Estrada, John, Sidike Paheding, Xiaoli Yang, and Quamar Niyaz. "Deep-learning-incorporated augmented reality application for engineering lab training." Applied Sciences 12, no. 10 (2022): 5159.

[20]. Mao, Chia-Chi, and Chien-Hsu Chen. "Augmented reality of 3D content application in common operational picture training system for army." International Journal of Human–Computer Interaction 37, no. 20 (2021): 1899-1915.

[21]. Binsch, Olaf, Charelle Bottenheft, Annemarie Landman, Linsey Roijendijk, and Eric HGJM Vermetten. "Testing the applicability of a virtual reality simulation platform for stress training of first responders." Military Psychology 33, no. 3 (2021): 182-196.

[22]. Malik, Ali Ahmad, Tariq Masood, and Arne Bilberg. "Virtual reality in manufacturing: immersive and collaborative artificial-reality in design of human-robot workspace." International Journal of Computer Integrated Manufacturing 33, no. 1 (2020): 22-37.

[23]. Park, Kyeong-Beom, Minseok Kim, Sung Ho Choi, and Jae Yeol Lee. "Deep learning-based smart task assistance in wearable augmented reality." Robotics and Computer-Integrated Manufacturing 63 (2020): 101887.

[24]. Boletsis, Costas, and Jarl Erik Cedergren. "VR locomotion in the new era of virtual reality: an empirical comparison of prevalent techniques." Advances in Human-Computer Interaction 2019 (2019).

[25]. Nileshkumar R Patel, Shishir Kumar, Energy-Efficient Approach for Effective Estimation of Delimited Node Position with Limited References, The Computer Journal, Volume 61, Issue 6, June 2018, Pages 881–895, https://doi.org/10.1093/comjnl/bxx102

[26]. haka, G., Patel, N.R. (2016). Selecting Favorable Reference Nodes to Aid Localization in Wireless Sensor Networks. In: Proceedings of International Conference on ICT for Sustainable Development. Advances in Intelligent Systems and Computing, vol 409. Springer, Singapore. https://doi.org/10.1007/978-981-10-0135-2_63

[27]. Diwan, Tausif, G. Anirudh, and Jitendra V. Tembhurne. "Object detection using YOLO: Challenges, architectural successors, datasets and applications." multimedia Tools and Applications 82, no. 6 (2023): 9243-9275.

[28]. Bisht, R., Patel, N. (2023). An Approach for Effective Object Detection. In: Singh, M., Tyagi, V., Gupta, P., Flusser, J., Ören, T. (eds) Advances in Computing and Data Sciences. ICACDS 2023. Communications in Computer and Information Science, vol 1848. Springer, Cham. https://doi.org/10.1007/978-3-031-37940-6_3

[29]. https://www.kaggle.com/datasets/rtatman/ms-coco