Academic Science Journal

# Human Body Tracking Method Based on YOLOV5s Object Detection

**Shaymaa Tarkan Abdullah, Bashar Talib AL-Nuaimi and Hazim Noman Abed**

Department of Computer Science – College of Science – University of Diyala

scicompms2114@uodiyala.edu.iq

**Received: 20 September 2022** **Accepted: 21 November 2022**

DOI: https://doi.org/10.24237/ASJ.01.04.692C

## Abstract

Body tracking is a viable solution for interacting with the human-computer and augmented reality. They are considered a necessity for a comprehensive understanding of human mobility. According to a robust tracking system of the human body, locating and tracing the human body in practical applications are challenging due to the enormous number of deformations and variations in body parts, postures, skin colures, lighting conditions, and clothes. To reduce complexity, several earlier works have concentrated on specific issues, such as face recognition, hand motion recognition, and mark recognition. The researchers agreed that the most previous approaches presume that people act while standing. Focusing on the detection and tracking of the human body, and the other component focusing on the detection and analysis of human action, the suggested system is based on the integration of tracking with the study of human activities through movement. And use YOLOV5s algorithm for detection and tracking , the result achieve by this algorithm and proposed system mAp 99%. That is replacing typical processing procedures like roboflow with an alternative that can function without the Internet. The proposed system is based on ten classes that contain a collection of overlapping verbs and can operate on photos, videos, and real-time systems.

**Keywords:** Body Tracking, Human Action Detection, YOLOV5sAlgorithm, Standfor40 Dataset.

Academic Science Journal

# طريقة تتبع جسم الإنسان على أساس خوارزمية يولو اكتشاف كائن

**شيماء تركان عبد الله، بشار طالب حميد وحازم نومان عبد**

قسم الحاسبات – كلية العلوم – جامعة ديالى

## الخلاصة

يعد تتبع الجسم حلاً قابلاً للتطبيق للتفاعل مع الكمبيوتر البشري والواقع المعزز. تعتبر ضرورة لفهم شامل للتنقل البشري. وفقًا لنظام تتبع قوي لجسم الإنسان فإن تحديد موقع جسم الإنسان وتعقبه في التطبيقات العملية يمثل تحديًا نظرًا للعدد الهائل من التشوهات والاختلافات في أجزاء الجسم، المواقف، ألوان البشرة، ظروف الإضاءة والملابس. لتقليل التعقيد ركزت العديد من الأعمال السابقة على قضايا محددة مثل التعرف على الوجوه، التعرف على حركة اليد والتعرف على العلامات. اتفق الباحثون على أن معظم الأساليب السابقة تفترض أن الناس يتصرفون أثناء الوقوف مع التركيز على كشف وتتبع جسم الإنسان والمكون الآخر الذي يركز على كشف وتحليل عمل الإنسان. استند النظام المقترح إلى تكامل التتبع مع دراسة الأنشطة البشرية من خلال الحركة. مع استخدام YOLOV5s تم استبدال المعالجات النموذجية بنظام يعمل بدون انترنيت. يعتمد النظام المقترح على عشرة أصناف تحتوي على مجموعة من الأفعال المتداخلة ويمكن أن تعمل على الصور ومقاطع الفيديو وأنظمة الوقت الفعلي.

**كلمات مفتاحية:** تتبع الجسم، الإنسان، الكشف، الأنشطة، الفئات.

## Introduction

Vision-based tracking of human bodies has been a critical research topic over the past decade due to its potential applications in areas such as human-computer interaction (HCI), surveillance, and motion capture [1]. It poses several challenges for motion tracking technology, including high dimensionality, shifting human forms, and complex dynamics. In recent years, there has been a close connection between tracking and activity identification, which has been shown to play a significant part in assessing individuals' behavior based on their actions [2]. Recognizing human activity, sometimes known as HAR, is one of computer vision's that is the most essential and challenging topics. Gaming, human-robot interaction, rehabilitation, sports, health monitoring, video surveillance, and robotics are just a few fields that could benefit significantly from its implementation [3]. Since the beginning of the field of computer vision, action recognition has been one of its most important goals, and it has made

tremendous strides forward in recent years [4]. It is easy to fall into the trap of thinking that identifying human activity is a straightforward task. There are issues with scenarios and advanced movement that have a high pace. The application of artificial intelligence (AI) for activity prediction based on numerical analysis has captured the interest of a significant number of researchers. To modify the comparison of these methods, numerous datasets on tagged acts that are very distinct from one another it terms of their content, and methodology have been generated [5]. The activities of humans create significant roadblocks in a variety of industries. There are several user-friendly applications in this area, including intelligent homes, valuable artificial intelligence, human-computer interactions, and enhancements in protection in many areas, such as security, transportation, education, and medication through the management of falls, or assisting the elderly with medication consumption [6]. Smart homes are one of the user-friendly applications in this area. Citation needs the development and success of deep learning techniques in various computer vision applications lend support to the idea that these techniques could be used in video processing. When doing an activity-based analysis of human behavior, humans present a considerable challenge. It is possible to depict more than one people in video sequence by their bodies' motion, skeleton, and abstraction elements [7].

**Objective of paper**

The suggested system works to demonstrate the optimal number of objects for detection and tracking systems, where ten items with varying outcomes for each model where the model was worked on, so resolving the issue of overlapping things and failing to accurately discriminate between them. In the first case, we have 10 in one modal, but in the second, we have two entities in one modal.

**Challenge of Tracking with Action Detection**

One must determine a person's kinetic states to recognize human actions effectively. Walking and running are daily human activities and are easy to distinguish. Complex tasks like "peeling an apple" are harder to identify. difficult tasks can be divided into more specific and easier-to-recognize activities. Detecting things in a scene can help humans better understand ongoing

events [8]. Frontiers most human activity recognition work requires a figure-centric location with an uncluttered background. Background clutter, partial occlusion, variations in scale, viewpoint, lighting and appearance, and frame resolution make developing a completely automated human activity detection system difficult. Annotating behavioral role takes time and event information. Intraclass and interclass similarities complicate the problem [9]. Different bodily movements express actions within the same class, while activities within classes may be hard to discern due to comparable information. Humans' habits make detecting the underlying action difficult. Inadequate benchmark datasets make it difficult to build a real-time visual model for learning and understanding human movements. To solve these challenges, you must do three things:

• background subtraction: the system separates invariant image components (background) from moving or changing objects (foreground).

• Tracking human mobility over time

• human action and object detection, which localizes human activity in a picture. Human activity recognition examines video or still images for activities. They use this knowledge to identify input data accurately.

Figure (1) illustrates show the backdrop impacts detection.

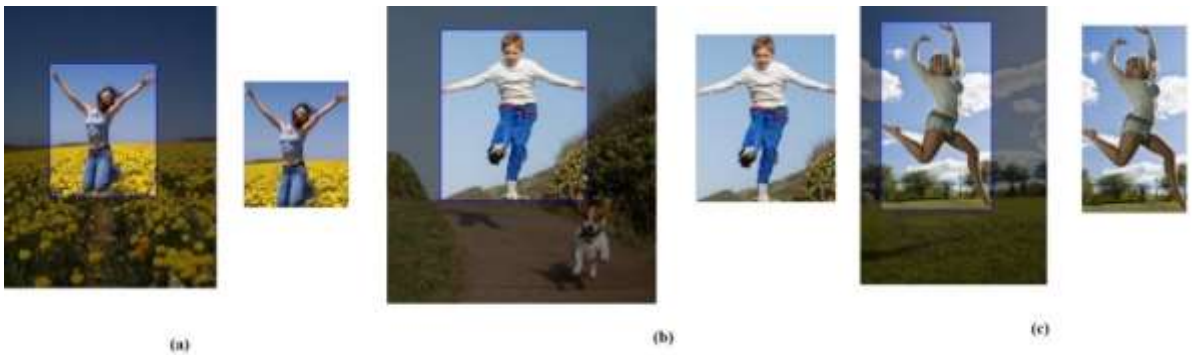


**Figure 1:** Background effect on action detection

Since the background plays an essential and vital role in Figure (1), there are three models of jumping movement. Still, once the object, i.e., the human being, is identified and separated from the background, the signal can overlap between walking and not jumping in (a, b). It is possible to dance Balinese instead of jumping in figure (C).

**Related Work**

Many studies in the field of human tracking and action have been published in recent years, and this paper highlights a few of them:

**L.Liu *et al.* (2019)** [10]**:** proposed a system for Idiosyncratic circumstances for action recognition in still photos that can be overcome using a multi-task learning approach. To suppress the activation of deceptive objects or backdrops, route the network's activations such that it concentrates on humans. Human-mask loss automatically activates feature maps based on the target human's face. Offer a multi task deep learning system that simultaneously predicts the human action class and the human location heatmap. This approach produces mAP results scoring 94.06 percent on the Stanford40 dataset and 40.65 percent on the MPII dataset. And the limitation of the system is. It entails merging human and object interaction strategies to better leverage action-relevant circumstances in the images supplied, and the proposed method works only on Still Images.

**Sattar Chan *et al.* (2019)** [11]**:** Three networks are used to determine human posture: the most relatable object in the scene, and the overall context, including actors and things around the person being evaluated. Before testing the suggested method, the conventional transfer learning method is assessed using four standard pre-trained convolutional neural networks for features extraction and Support vector machine classification. Only the SVM's primary components are used to predict human action. The model evaluation uses the Stanford40 dataset. This collection includes 40 images of human activities, each with a bounding box. There are 9532 photographs, with 180-300 picture assigned to each class; however, only ten dataset categories are used for the experiment, and the proposed system results mAP 87.1%. Transfer learning avoids the

computationally intensive and time-consuming process of training a deep learning model from scratch.

**A. Raza *et al.* (2020)** [12]**:** A pre-trained Convolutional Neural Network (CNN) model is used as a feature extractor, followed by a Support Vector Machine (SVM) classifier for action recognition. CNN information from an extensive data set can be transferred to problems with minimal training data. The suggested method is assessed on the stanford40 human action data set, which includes 40 kinds of activities and 9532 photos. The proposed method extracts deep representations from Resnet-18's last pooling layer, pool5. It achieves 87.22% accuracy on the dataset using deep representations and a state-of-the-art SVM classifier. The technique only works with still photos, not videos or front-and-back photographs.

**A. Diba *et al.* (2021)** [13]**:** The proposed system for human activity recognition uses a pre-trained CNN model as a feature extractor and a Support Vector Machine (SVM) classifier for action recognition. Previous CNN knowledge from an extensive data collection can be applied to activity identification tasks with less training data. The suggested method is assessed on the stanford40 human action data set, which includes 40 kinds of activities and 9532 photos. The proposed method extracts deep representations from the last pooling layer and pool5 in Resnet-18. Following these deep representations, a state-of-the-art SVM classifier predicts the action class in a given images; the suggested method achieves 87.22% accuracy on the dataset. The technique only works on static photographs, not video or front and back shots.

**S. Mohammadi *et al.* (2021)** [13]**:** Using pre-trained CNNs, they utilize transfer learning to address the shortage of big action recognition datasets with labels. In addition, because the final layer of the CNN contains class-specific information, they apply an attention method to the CNN's output feature maps to extract more discriminative and robust features for categorizing human behaviors. In addition, our methodology employs eight distinct pre-trained CNNs and evaluates their performance on the Stanford 40 dataset. Lastly, they propose using Ensemble Learning to improve the classification accuracy of actions by pooling the predictions of multiple models.

The optimal results can achieve 93.17 percent accuracy on the Stanford 40 dataset. The system's limitation is Focusing only on still images and not taking into account the moving footage or videos because it is assumed that the still it is the most difficult.

**K. Hirooka *et al.*(2022)** [14]. Multi-channel attention networks with transfer learning suggested a convolutional neural architecture. This study uses four CNN branches to assemble features. An attention module features extracted from pre-trained models' feature maps in each branch. Finally, the four branches' feature maps were concatenated and submitted to a network for national recognition. they tested your system with the Stanford 40, BU-101, and Willow datasets. Stanford 40 has 93.76% accuracy, BU-101 97.98%, and Willow 92.44%.

## Material and methods

**proposed system**

The object's discovery depends on how the person interacts with the things in the picture and recognizes what the person does through these objects. In the proposed system, work was doing based on the person's analysis of what he does through he carries or deals with and research about the action, he performs and the proposed system. It works on the other side, which is also tracking through videos, where people are tracked, and their efforts analyzed, meaning that the system can work offline and online, i.e., in real-time. Figure (2) is the general outline of the proposed method.

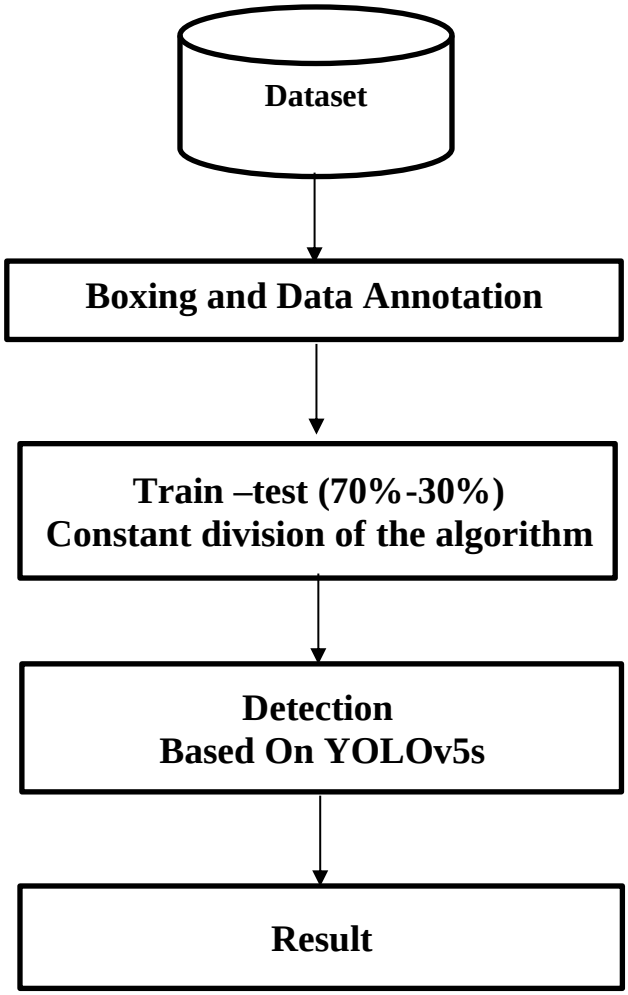


**Figure 2:** Block Diagram of Proposed System

**Dataset**

People are shown participating in various pursuits across various settings in the photographs that make up the Stanford 40 Action Dataset. Provide a bounding box of the human in each shot, indicating that the person is engaged in the activity mentioned in the image's filename. There are 9532 photos total, each class with 180 and 300 images, respectively. For instance, brushing one's teeth, sweeping the floor, reading a book, and throwing a Frisbee; figures illustrating examples of photos taken from this dataset can be found in Figure (3).

Academic Science Journal


**Figure 3:** Example images from Stanford 40 dataset

Figure 4 shows the class of action and number used in the proposed system.

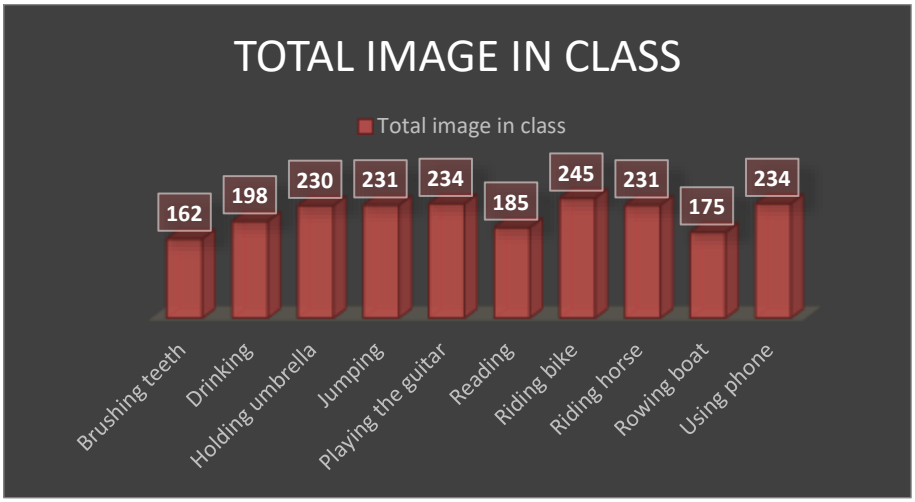


**Figure 4:** Preparing Images in Each Class

The following Figure Shows the volume of training for the test.

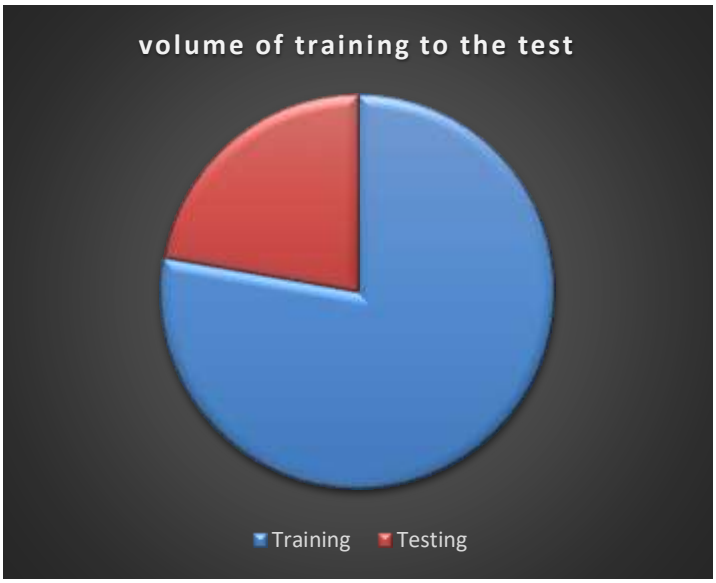


**Figure 5:** Volume of Training to The Test.

**Image pre-processing**

They are primary processing operations that occur on the image to improve the it and thus improve the system's accuracy. The following steps are the processes used in the proposed method.

**Boxing and Annotation**

The YOLOv5s Algorithm must be trained on a dataset that includes images containing a label (Data Annotation) in which the coordinates of the bounding squares are mentioned. The annotations used with the YOLOv5s algorithm must be with an extension TXT; Most of the methods used in the naming process are with limited efficiency, and the most famous method is to use of Roboflow software that supports object detection and classification models. and it requires either a paid subscription to get the full features or for free with specific features, and it has a slow implementation for two; also, a key factor that is a program, that works only online and does not work offline. From this standpoint, a sub-program was made, and they were

included within the proposed system that works on the box and data annotation for each image and stored in an array in files with the extended text. The following points decryption the program design to make boxing and annotation for ten classes in figures form to Clarify how it works for classes.

In the proposed system to making annotation and boxing, two options:

- edges
- Grayscale

Where the benefit of these two options if dealing with high-noise, low-resolution images with contrast in brightness, it is possible to use these two characteristics to make the Annotation process. After the box application phase and annotation work, this resulting image is stored in an array in a txt file to be handled by the YOLOV5s Algorithm.

**2.4 Architecture of YOLOv5s**

The structure of YOLOv5s is divided into four parts: Input, Backbone, Neck, and Prediction (Head).

**Input:** Adaptive image scaling occurs in input components; In generally used target detection methods, various images have different lengths and widths. Therefore, the standard way is to scale the source images to a standard size and then feed them to the detection network.

**Backbone:** Backbone is a convolutional neural network that accumulates and produces image features, as its feature backbone is designed to eliminate picture properties. In the backbone, the operation focus structure occurs; the construction of YOLOv5s, an original (256 256 3 images), is fed into the Focus structure, and then the slicing operation is used to convert that to a (128 128 12) feature map, followed by 32 convolution operation kernels that produce a (128 128 32) final feature map.

**Neck:** A sequence of network layers mixes and integrates visual features, with the neck's function being to collect feature mappings from different stages and transport them to the

prediction layer. The PLANET network mainly utilizes the neck to build feature pyramids. that enable the suggested approach to generalize object scaling successfully. The feature pyramid improves the proposed system's detection of objects at various scales, allowing it to recognize the same object at multiple sizes and scales. A set of network layers that integrate and mix picture features can be utilized to increase the diversity and resilience of features, and move the picture characteristics to the prediction layer.

**Head:** The proposed system Head is mainly used for the final detection, predicting image features, generating bounding boxes and predicting categories. It applies anchor boxes on feature maps and generates final output vectors with class probabilities, object scores, and bounding boxes.

## Evaluation of result

There are a variety of evaluative measures that can be employed based on the activities. When analyzing the findings of a new method or comparing them to those of other current systems, it is vital to use the same datasets to assure compatibility. The proposed system should use Precision, Recall, Mean Average Precision (mAP), and F1 to analyze the results. Several experiments were undertaken, and the subsequent experiment synthesized all the classes' efforts. It focused on the first, the movement of riding a bicycle, using a telephone while holding an umbrella, playing the guitar at home, brushing one's teeth, kayaking, reading on a boat, drinking, and jumping stage being to complete the job box. Also, the Annotation displayed below (6), an example of a scanning technique, is shown below.

Academic Science Journal

**Figure 6:** Example of riding the bike, using the phone and holding umbrella, riding house, playing the guitar, brushing teeth and rowing boat, reading, drinking and jumping.

The result of the detection class in Yolo v5s show in Figure (7):

**Figure 7:** the result of detection-based Yolov5s.

Confusion matrix measures were used to evaluate the results obtained for the test part, which are shown in Figure(8). Based on the above Figure, the values of the matrix are as follows:

Academic Science Journal

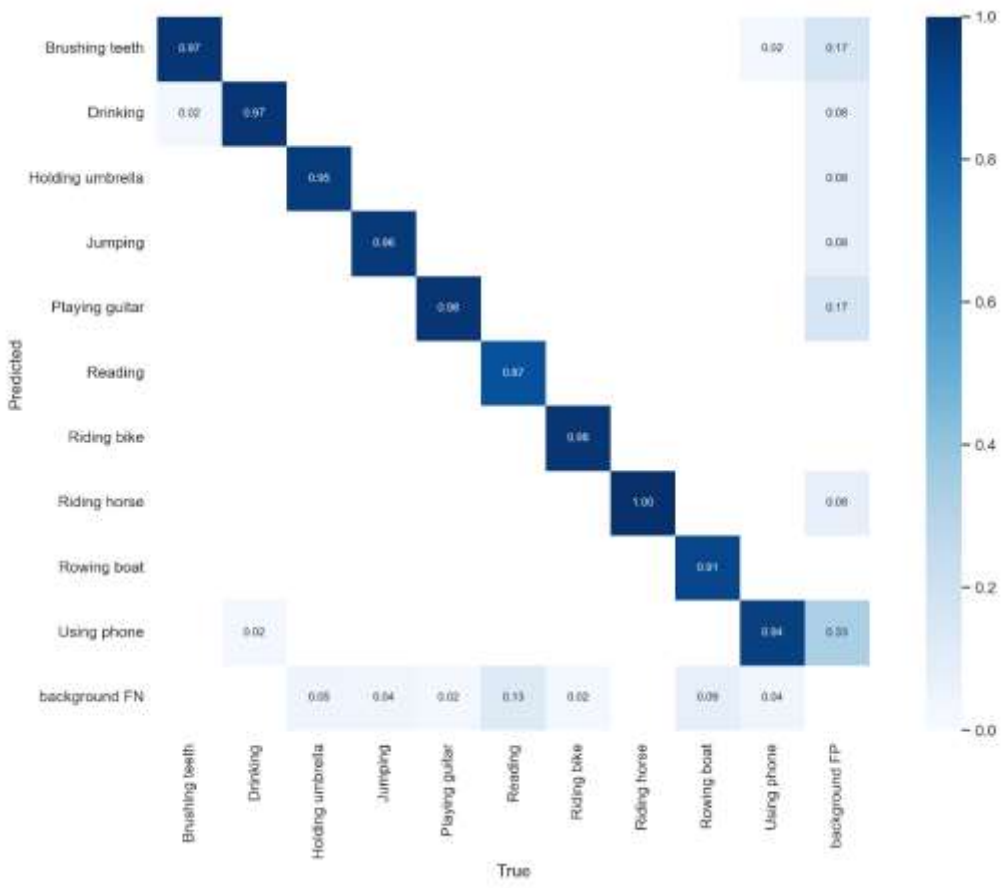


**Figure 8:** Result of Confusion Matrix.

Based on the above figure, the values of the matrix are as follows:

**Table 1:** Values of The Confusion Matrix

| Parameter | riding bike | Using phone | holding umbrally | riding horse | play guitar | brushing teeth | rowing boat | reading | drinking | jumping |
|---|---|---|---|---|---|---|---|---|---|---|
| TP | 0.98% | 0.94% | 0.95% | 1.00% | 0.98% | 0.97% | 0.91% | 0.87% | 0.97% | 0.96% |
| FP | 0 | 0.33% | 0.08% | 0.08% | 0.17% | 0.17% | 0 | 0 | 0.08% | 0.08% |
| FN | 0.02% | 0.04% | 0.05% | 0 | 0.02% | 0 | 0.09% | 0.13% | 0 | 0.04% |

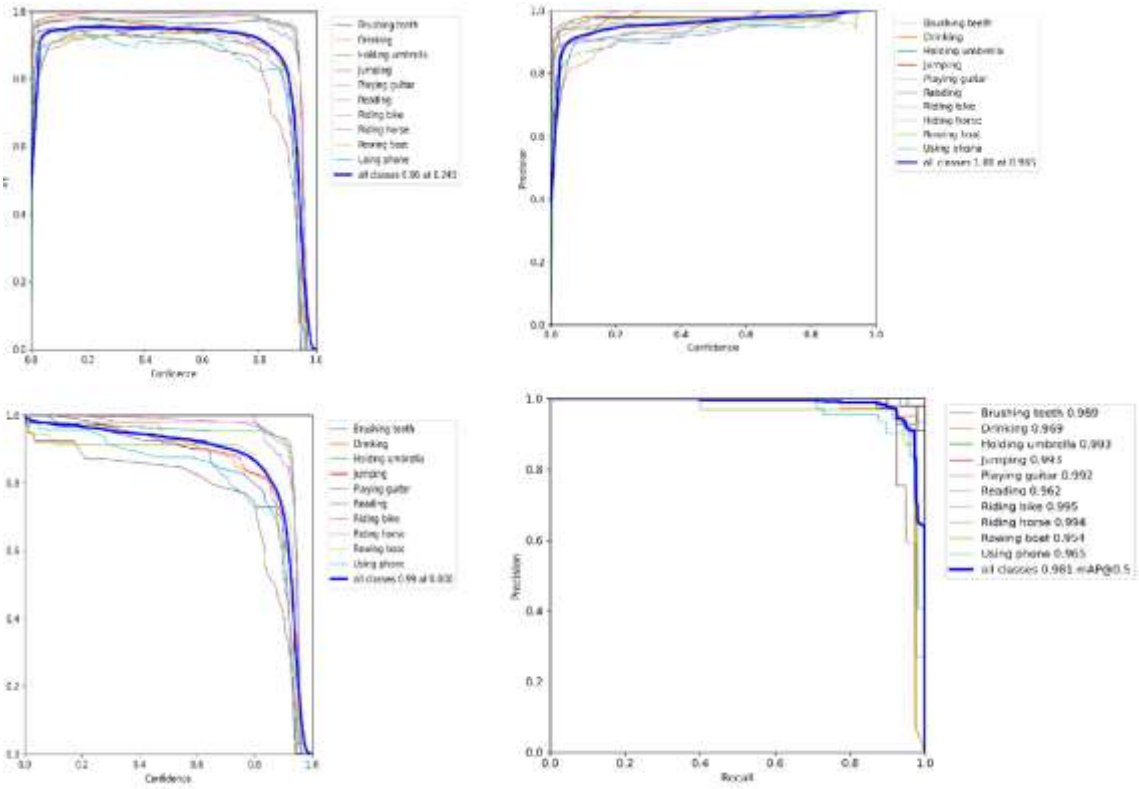

**Figure 9:** All Results of Experiment four

The following table shows the result of ten experiments.

**Table 2:** Evaluation Result of Experiment ten

| Scale | Result |
|---|---|
| F1 | 0.98% |
| Recall | 1.00% |
| Precision | 1.00% |
| mAP | 0.99% |

**Detection and Tracking in Vide Real-Time**

The work has the advantage of tracking and detecting natural and real-time videos; thus, the system can see and trace human being. The following is an experiment that demonstrates this work in the videos.

Academic Science Journal

**Figure \10:** Example of Tracking and Detection in Video-Real Time.

**Comparison with Previse Studies**

The following table is a comparison of the proposed system with the previous works, presented in summary form for comparison in terms of accuracy

**Table 3:** Comparison with Previse Studies

| RF. | Dataset | Methods | Result |
|---|---|---|---|
| 2019 [10] | • Stanford40<br>• MPII | multi-task CNN | mAP<br>Stanford40<br>94.06%<br>MPII<br>40.65% |
| 2019 [11] | Stanford40 | 1. CNN resnet18<br>2. SVM | mAP<br>Stanford40<br>87.1% |
| 2020 [12] | Stanford 40 | Resnet-18<br>svm | Accuracy<br>87.22% |

| | | | |
|---|---|---|---|
| 2020<br>[1] | Stanford 40 | CNN | accuracy<br>93.17 % |
| 2021<br>[13] | 3. PASCAL VOC 2012<br>4. Stanford 40<br>5. Berkeley dataset | CNN | mAP<br>6. PASCAL VOC 2012<br>75.4&<br>7. Stanford 40<br>77.6%<br>8. Berkeley<br>86.6% |
| 02022 | 9. Stanford 40<br>10. BU-101<br>11. Willow | CNN | 12. Stanford 40<br>93.76% accuracy<br>13. BU-101<br>mAP<br>97.98%<br>14. Willow<br>mAP<br>92.44% |
| Proposed system | 15. Stanford 40 | YOLOv5s | F1<br>0.98%<br>Recall<br>1.00%<br>Precision<br>1.00%<br>mAP<br>0.99% |

The following table is a statistical chart that shows the difference between the proposed system and the previous studies in terms of accuracy, knowing that the system has a new idea, as ten objects were worked on in an integrated manner. They did not address this but instead worked on the principle of each class containing one thing.

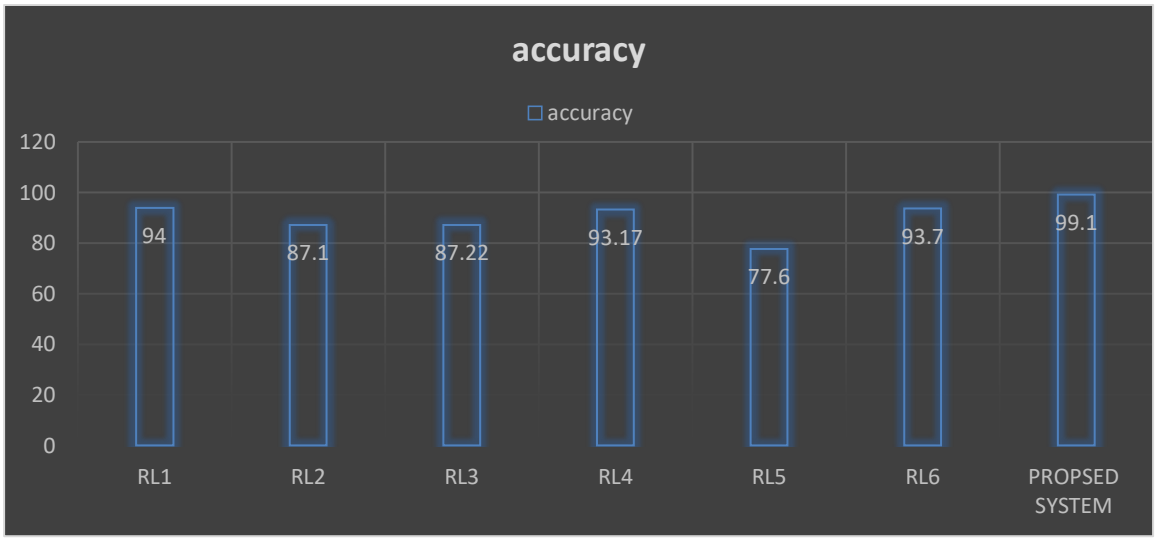


**Figure 11:** The Difference Between the Proposed System and The Previous Studies

## Conclusions

The suggested system tracks people and analyzes their actions using the yolyv5s algorithm, which has excellent tracking and detecting capabilities. The processing actions employes in tracking or detection are through Internet programs like Robflow, which operates as a box and annotation to identify the object and produce a txt formula for the algorithm to deal with files through annotation. As a result, most apps must offer the Internet in two versions (free and paid) (not free). A program in the system executes this process offline and with the same result without the internet. Also, human behaviour has been studied in many ways. The suggested system detects ten actions for each movement. The effort is on ten integrated activities, where the design incorporates the results of ten experiments, i.e., ten classes. The first and second classes contain the first class's activity with additions and up to ten courses of combined action. The proposed system can function offline on photos and videos and in real-time to track and analyze human movements.

## References


1. S. Mohammadi, S. G. Majelan, S. B. Shokouhi, Ensembles of deep neural networks for action recognition in still images, In: 2019 9th Int. Conf. Comput. Knowl. Eng. ICCKE 2019, pp. 315–318(2020)
2. M. Webber, R. F. Rojas, IEEE Sens. J., 21(15), 16979–16989(2021)
3. S. K. Yadav, K. Tiwari, H. M. Pandey, S. A. Akbar, Knowledge-Based Syst., 223, (2021)
4. Y. Huang, T. S. Huang, Proc. - Int. Conf. Pattern Recognit., 16(1), 552–555(2002)
5. M. Vrigkas, C. Nikou, I. A. Kakadiaris, Front. Robot. AI, 2(NOV), 1–28(2015)
6. A. W. Muhamada, A. A. Mohammed, ADCAIJ Adv. Distrib. Comput. Artif. Intell. J., 10, (4), 361–379(2022)
7. V. T. Le, K. Tran-Trung, V. T. Hoang, Comput. Intell. Neurosci., vol. 2022, (2022)
8. F. Gu, M. H. Chung, M. Chignell, S. Valaee, B. Zhou, X. Liu, ACM Comput. Surv., 54(8), (2022)
9. D. R. Beddiar, B. Nini, M. Sabokrou, A. Hadid, Multimed. Tools Appl., 79(41–42), 30509–30555(2020)
10. L. Liu, R. T. Tan, S. You, Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), 11365 LNCS, 152–167(2019)
11. A. S. Chan, IJCSNS Int. J. Comput. Sci. Netw. Secur., 19(11), (2019)
12. A. R. Siyal, Int. J. Adv. Comput. Sci. Appl., 11(5), 471–477(2020)
13. A. Diba, A. Mohammad Pazandeh, H. Pirsiavash, L. Van Gool, K. Leuven, IEEE Comput. Intell. Mag., 3(4), (2021)
14. K. Hirooka, M. A. M. Hasan, J. Shin, A. Y. Srizon, IEEE Access, 10, 47051–47062(2022)