Ministry of Higher Education
and Scientific Research
University of Diyala
College of Science
Department of Computer
Science

Human Activities Detection method Using YOLOv5s Algorithm

A Thesis

Submitted to the Department of Computer Science\ College of Science\ University of Diyala in a Partial Fulfillment of the Requirements for the Degree of Master in Computer Science

By

Shaymaa Tarkan Abdullah

Supervised By

Assist. Prof. Dr. Bashar. Talib

Assist. Prof. Hazim Noman Abed

2022 A.D. *IRAQ\ Diyala* *1444 A.H*

بسم الله الرحمن الرحيم

﴿ يَرْفَعِ اللَّهُ الَّذِينَ آمَنُوا مِنكُمْ وَالَّذِينَ أُوتُوا
الْعِلْمَ دَرَجَاتٍ وَاللَّهُ بِمَا تَعْمَلُونَ خَبِيرٌ ﴾

سورة المجادلة آية ١١

Dedication

To the flower of life and its light and the most precious person in my life, my tender mother
To whom I proudly carry your name, teach me how to make successful and instill confidence in myself my dear father, may God extend your life.
To those who supported me in adversity and the source of my happiness, my companion to my path, and my love, my dear husband Laith.
To whom their love and blood flow in my veins, and I lived with them the most beautiful moments, my sisters and brother .
To the candles that illuminate the path of knowledge in my path, my distinguished teachers.
To everyone who loved me and supported me in my scientific and practical life.

Shaymaa Tarkan

Acknowledgment

Initially, I extend my sincere thanks to "Allah" who did not abandon me and helped me complete this work

Many thanks are to my supervisors, " Asst. Prof. D.r Bashar Talib AL-Nuaimi " And " Asst. Prof.Dr. Hazim Noman Abed "for their contribution, support, and discussions which help me a lot during the research period.

Though only my name appears on the cover of this thesis, many people have contributed to its production.

I will never forget to thank the current staff of the Department of Computer Science\ College of the Science\ University of Diyala in a partial fulfillment of the Requirements for the master in computer science for their various forms of support during my master's study. On top of them is the head of the department.

Abstract

Human activities recognition has been a highlighted topic for researchers because of its various applications which include video surveillance, Human- machine interaction, ambient-assisted living, smart system design and autonomous driving. Earlier works focused on facial, hand motion, and mark recognition to decrease complexity. However, human activities recognition systems have many problems one of these tracking systems based on the activity of the human body do not give an analysis of what the person does but only the work of tracking the human body.

In this thesis system has been introduced for human body tracking and activity recognition based on deep learning using YOLOv5s algorithm. The aim is to design system that has been functional by using a certified global dataset (stanford40 dataset) based on ten classes that contain a collection of overlapping actions and can operate on photos, videos, and real-time systems.

The results achieved by the proposed system were (mAP: 99 %, Precision: 100%, Recall: 99%, F1-score: 99%).

List of Contents

List of Figures

List of Tables

List of Algorithms

List of Abbreviations

Abbreviations	Meaning
AI	Artificial Intelligence
AP	Average precision
CNN	Convolutional Neural Network
DL	Deep Learning
FN	False Negative
FP	False Positive
HAR	Human Activity Recognition
HCI	Human-computer interaction
IOU	Intersection Over Union
mAP	Mean Average Precision
ML	Machine learning
SSD	Single-Short Detector
SVM	Support-Vector machines
RGB	Red, Green, And Blue
TN	True Negative
TP	True Positive
YOLO	You Look Only Once

Chapter One
General Introduction

Chapter one

General Introduction

1.1 Overview

Human body tracking based on computer vision has been a major study subject in the last decade due to its potential uses in surveillance, motion capture, and Human-Computer Interaction (HCI). It presents several technical problems for motion tracking, including high dimensionality, changing human forms, and complicated dynamics. In modern times, there is a strong link between tracking and activity detection, and it plays a vital part in assessing people's behavior based on their activities [1].

Human activity identification is one of computer vision's most essential and difficult topics. It is essential for a variety of functions, including "gaming," "human-robot interaction, rehabilitation, sports, health monitoring, and video surveillance [2].

Since the advent of computer vision, action recognition has been a crucial goal, and recent years have seen substantial advancements. Sometimes, the identification of human behavior is seen as a straightforward procedure. There are problems in advanced scenes with high velocity.

Various datasets on tagged acts with substantial differences in content and approach were developed to change the comparison of these methods. Human actions present considerable obstacles in numerous sectors. Intelligent homes, usable artificial intelligence, human-computer interactions, and enhanced security in a range of fields are just a few of the user-friendly applications in this area. Protection, travel, education, and health care are all areas that need improvement (via the management of falls or assisting the elderly with medication consumption) [3].

The use of deep learning methods in video processing is supported by their development and success in a numbers of computer vision applications. The presence of people presents a considerable challenge to the activity-based study of human behavior. Movement, skeleton, and abstraction are all ways that people might be shown in moving images [4].

1.2 Body Tracking

Motion capture for human animation, human-computer interface, interactive virtual worlds, and video surveillance are just a few fields where human body tracking is a hot topic for study [5]. To improve the human-computer interface, humans and machines must find the same perceptual aspects fascinating. The utilization of human body parts, such as facial and kinetic gestures, is increasingly integral to human-computer interaction [6].

Incorporating body tracking with HCI has shown to be a fruitful combination. Human motion recognition systems now widely acknowledge that body tracking is a crucial condition for achieving good results. Recognizing and keeping track of a human being is challenging because of the wide variety of people, poses, skin tones, lighting conditions, and clothes that might be encountered. Facial tracking, hand motion tracking, and marker tracking are only a few examples of the specialized approaches that have been used in the past to tackle this problem. Many efforts have been made to detect, monitor, and understand human activities, with the prevailing assumption being that people do these actions while standing still [5].

It is impossible for a computer to handle all the issues that might come up from interacting with a person. If the user is only able to see the currently focused area of the screen, the computer's processing load for the program may be reduced, allowing for a smoother real-time experience [7]. Figure (1.1) shows an example of body tracking.

Figure (1.1): Example of Human body tracking [8].

1.3 Human Action Recognition

Human action recognition is vital for interpersonal communication. It's hard to extract because it reveals a person's identity, mentality, and mental state. Diverse activity recognition systems are required for a wide range of uses, including but not limited to video surveillance, human-computer interaction, and robots describing human behavior. Although difficult, human action recognition is essential for categorizing actions in videos and analyzing them. It is being installed as a component of the ongoing monitoring of human behavior [9].

It has also been suggested for use in areas such as eldercare and law enforcement, sports injury prevention, location estimation, and home surveillance. Even though great progress has been achieved in human action identification from surveillance videos, this shortcoming remains owing to variables such as perspective and camera distance variations, the context's personality, and speed variance.

Many techniques for detecting a moving individual employ background extraction. Several strategies, in addition to Gaussian distribution, include linking human behavior using various motion tracking techniques. Human monitoring is undertaken to track its motion and generate trajectories for all possible sequences [10]. The following Figure (1.2) shows an example of human action recognition.

Figure (1.2): Examples of Human Action Recognition [9].

Human activities can be broken down into the following categories, according to the degree of difficulty they present:

- Gestures.
- Atomic actions.
- Human-object interactions and human connections.
- Group actions.
- Behaviors.
- Events.

The division of human activities according to their complexity is shown in Figure (1.3).

Figure (1.3): Deconstruction of Human Activities [9].

1.4 Human Activity Recognition

Standard machine learning techniques can be used in human activity recognition. However, standard machine learning algorithms for Human Activity Recognition (HAR) require the creation and selection of relevant characteristics. This requires painstaking human interaction and specialized expertise, and the developed and chosen characteristics may still produce unsatisfactory performance [11].

When describing a particular human action that has been accomplished, it is common to have trouble choosing the appropriate phrase. Typically, candidates use the phrases "action," "activity, and "behavior. Despite their apparent similarities, according to the Oxford English Dictionary, Action refers to anything that is done; activity refers to the state of being actively engaged; and behavior refers to how one conducts oneself in the context of one's external interactions in life. Each has its own meaning and can generate distinct jargon-based expressions [2] , table (1.1) refer to the main type of human activities .

In the fields of philosophy and sociology, these words may be described in more depth. Max Weber discussed the connection between activity and social action. consider an actor's actions to be purposeful and emotionally or experientially motivated. However, behavior is nothing more than a reflection of signals or impulses. By citing Weber's work, Campbell acknowledges that action is a purposeful activity, while the behavior is an activity conducted without purpose or intent [12].

Table (1.1): Main Categories of Activities[13]**.**

No.	Category	Activities
1	Locomotion	Walking, running, jogging, lying, standing, sitting, ascending and descending stairs are all examples of physical activity.
2	Transport mode	Transportation modes include bicycling, taking a bus, driving, and vehicle travel.
3	Phone usage	Texting, calling, using an app, surfing the web, and checking email are examples of common mobile phone activities.
4	Entertainment	Participate in soccer, basketball, social events, and gaming.
5	Health-related activity	Falls, respiration, rehabilitative activities, and cigarette smoking
6	Daily activity	Napping, using a laptop, traveling, eating, socializing, having a discussion, and going to work are examples of daily activities.
7	Gesture	Gestures of the body, arms, hands, and head, as well as body and sign language
8	Emotion	Positive, negative, fear, happiness, sadness, surprise, indifferent
9	Security	Presence, aggressiveness, and odd actions

The following Figure (1.4) summarizes the Applications of human activity recognition.

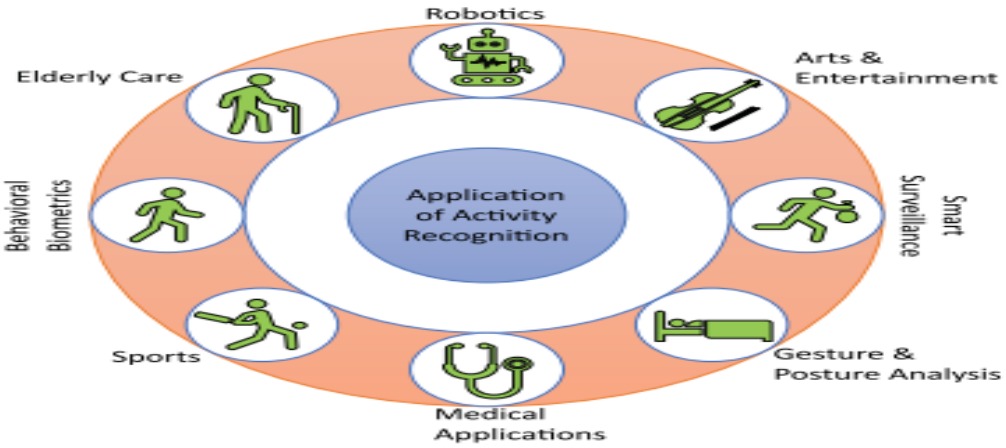

Figure (1.4): Applications of Human Activity Recognition [2].

1.5 Challenge of tracking and activity recognition

To train a computer to identify human actions, it is necessary to first ascertain the kinetic states of a person engaging in those actions. Human actions, such as "walking" and "running" are commonplace and simple to identify. On the other hand, it is more difficult to distinguish sophisticated tasks such as "peeling an apple" complex actions may be broken down into simpler activities, which are often easier to notice. Typically, spotting items in a scene may aid in the comprehension of human behaviors by giving useful information about the current occurrence [11].

The majority of human activity detection research has been done in situations where the actor is free to perform as they like without being constrained by the background because the figure is the main point. It is challenging to create a completely automated human activity detection system that can categorize a person's actions with minimal error due to factors such as backdrop clutter, partial occlusion, variances in size, perspective, lighting, appearance, and frame quality. Also, it takes a lot of work and context-specific data to annotate behavioral roles [12].

Variable people with different bodily motions convey actions within the same class, and it may be difficult to discern actions within classes since they may be represented by comparable data. The manner in which people conduct an activity relies on their habits, making it relatively difficult to pinpoint the underlying activity. In addition, it is difficult to develop a visual model for learning and understanding human motions in real time with insufficient benchmark datasets for assessment. In order to resolve these issues, a three part job is required [4], [9]:

- By using a process called "background subtraction" the system isolates the static components of a picture from the dynamic ones (foreground).
- Human tracking, whereby the technology identifies human movement over time
- human activity and object recognition.

This allows the algorithm to identify human activities inside a picture. Examining actions from still photos or video clips is the aim of human activity recognition. The following Figure (1.5) shows how the background affects detection action.

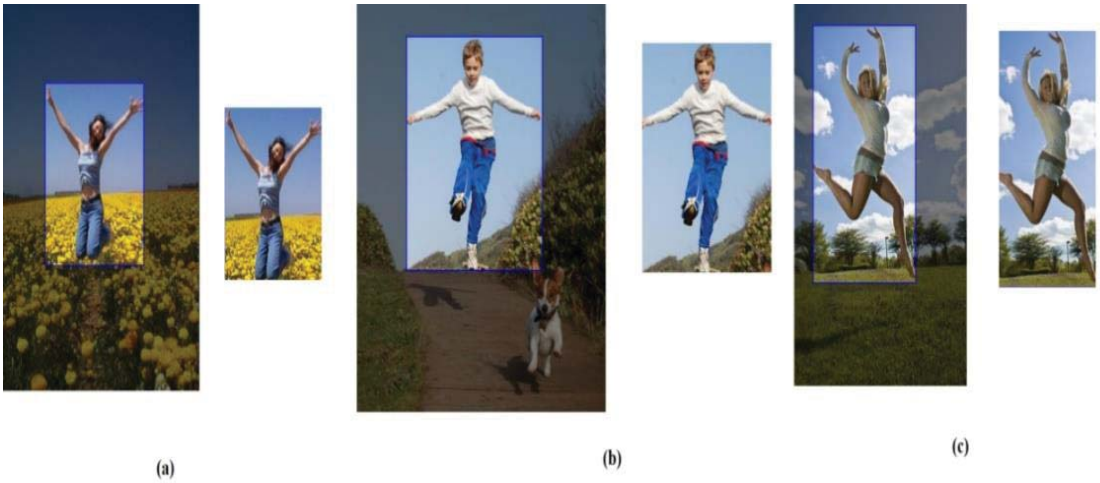

Figure (1.5): Background effect on action detection.

Since the background plays an essential and vital role depending on figure (1.5), there are three models of jumping movement. Still, once the object, i.e., the human being, is identified and separated from the background, the movement can overlap between walking and not jumping in (a, b). It is possible to dance Balinese instead of jumping in (c).

1.6 Related Work

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

1.L.Liu et al. (2019)[13]: They 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. Make available a deep learning system capable of doing simultaneous predictions on the human action class and the human location heatmap. Using this strategy, results were a mAP(Mean Average Precision) score of 94.06% on the Stanford40 dataset and a score of 40.65% on the MPII dataset. This is a limitation of the system itself. The proposed system is restricted to still images, and it needs a combination of human and object interaction strategies to make the most of the action-relevant scenarios captured in the provided photographs.

2. Sattar Chan et al. (2019)[14]: The suggested model has been provided, in which three networks are used to assess human posture, the most relatable item in the scene, and the entire scenario, which comprises actors and things associated with the person being examined. Four well-known pre-trained

convolutional neural networks are used for feature extraction, and the Support vector machine is used for classification, in order to assess the performance of the traditional transfer learning approach before assessing the performance of the proposed method. To anticipate human activity in the scenario, just the main components of the collected characteristics are fed through the SVM (support-vector machines). To assess the suggested model, the Stanford40 dataset is employed. This dataset contains photographs illustrating forty distinct human activities, with each picture including a bounding box of the subject doing the activity. There are a total of 9532 photographs, with 180-300 images allocated to each class; and the proposed system achieved a mAP of 87.1%. The computationally intensive and time-consuming nature of training a deep learning model from start may be circumvented through transfer learning.

3. *A. Raza et al. (2020)[15]:* They suggested to build an action recognition system that uses a well before Convolutional Neural Network (CNN) model to extract features and a Support Vector Machine (SVM) classifier to make final determinations. Evidence suggests that CNN expertise built up over a large dataset may be transferred to activity detection tasks that need less data for training. The suggested technique is assessed using the publicly accessible "stanford40 human action data set", which contains 40 kinds of activities and "9532 photographs". From the last pooling layer, deep representations may be derived using the architecture ("pool5 in the case of Resnet-18"). As a result of combining these deep representations with a state-of-the-art SVM classifier for action prediction, the suggested approach achieves 87.22% accuracy on the dataset. The system's major flaw is that it can't handle moving pictures, or even photos with different perspectives, such as a front and rear.

4. A.Diba et al. (2021) [16]: They proposed to implement a convolutional neural network capable of mining intermediate-level picture patches and

dedicated enough to tackling the relevant intricacies of the challenge. Specifically, educate as CNN named Deep Pattern, which was developed lately and is capable of learning discriminative patch groups. Two novel features make this an innovative effort overall. One, with our new method, crucial contextual information is taken into consideration. The pool of patches is then cleaned by a cyclic process of patch clustering and feature learning. They evaluate threir action categorization approach on the difficult "PASCALVOC 2012 Action and Stanford 40 Actions" datasets. For this, they use the Berkeley Attributes of dataset. The system's mAP results reveal that "PASCAL VOC 2012" achieves a success rate of 75.4%, Stanford 40 achieves a rate of 77.6%, and Berkeley achieves a rate of 86.6%. Qualitatively and quantitatively, the system's limitation shows promise, reaching state-of-the-art without relying on human posture or part annotations.

5. S. Mohammadi et al. (2021)[17]: They used 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 technique makes use of eight different CNNs that have already been pre-trained and assesses how well they perform using 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 setup 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 images are the most difficult.

6. K.Hirooka et al.(2022)[18]: They suggested a convolutional neural architecture by assembling multi-channel attention networks with transfer learning. For the purpose of performing feature fusion-based ensembling, this research made use of four different CNN branches. Each of the forks has an attention module built in for mining the feature map generated by the pre-trained models for contextual clues. At the conclusion of the day, the output for nationwide recognition was made by combining the feature maps acquired from the four distinct departments and then transmitting them to a network that was entirely interconnected. They assessed your system by employing three different datasets of human behaviors: the Stanford 40 actions, the BU-101, and the Willow dataset. The results of the system is that the Stanford40 has an accuracy of 93.76%, while the mAP for BU-101 is 97.98% and the mAP for Willow is 92.44%.

Table (1.2): summary of Related Work.

RF.	Dataset	Methods	Aim	Limitation	Result
2019 [13]	Stanford 40 MPII	multi-task CNN	Create a heat map of human activity types and proposed locations using deep learning.	proposed system works only on Still Images.	mAP Stanford40 94.06% MPII 40.65%
2019 [14]	Stanford 40	CNN resnet18 SVM	The proposed human action recognition method is proposed based on three networks utilizing transfer learning by pre-trained Convolutional neural	Because training a deep learning model from start is laborious and resource intensive, many researchers are turning to transfer	mAP Stanford40 87.1%

			network architecture and SVM classifiers in decision fusion where confidence scores of three different networks are related, and the final decision is produced	learning as an alternative.	
2020 [15]	Stanford 40	Resnet-18 svm	The proposed method, which is grounded in the transfer learning of deep representations, has been tested with a state-of-the-art SVM classifier trained on the target database and a pre-trained Resnet-18 convolutional neural network as the source architecture for feature extraction.	The system works only on the still image and does not work on video and images with front and still shots	accuracy 87.22%
2020 [16]	Stanford 40	CNN	propose using pre-trained CNNs to handle the lack of massive labelled data.	Focusing only on still images and not taking into account the moving footage or videos because it is assumed that the still images are the most difficult.	accuracy 93.17 %
2021 [17]	PASCAL VOC 2012 Stanford 40 Berkeley dataset	CNN	have dealt with human behavior and attribute categorization using discriminative visual cues of the moderate degree.	The qualitative and quantitative outcomes are encouraging, with state-of-the-art results being achieved without the need of human pose or part annotations.	mAP PASCAL VOC 2012 75.4& Stanford 40 77.6% Berkeley 86.6%
2022 [18]	Stanford 40 BU-101 Willow	CNN	presented in this research is a strategy that improves upon the accuracy with which human actions may be identified in still photographs.	concentrating on still pictures and ignoring moving pictures or movies in favor of just looking at still pictures is generally accepted that still, pictures are the most challenging to capture.	Stanford 40 93.76% accuracy BU-101 mAP 97.98% Willow mAP 92.44%

Based on what was mentioned in previous studies, most of the systems proposed by researchers only focused on the concept of a static image. They considered it one of the most challenging ways to work because it does not contain feelings, or it is possible to analyze the action or influences that facilitate the process of knowing the act. can work in images and videos together, and the main work close to the proposed system is [16] .

1.7 Problem Statement

Recognizing human activity in a scene is still a challenging and an important research area in the field of computer vision due to its various possible implementations on many fields including autonomous driving, bio medical, machine intelligent vision etc. But human activities recognition have many problem one of these the problem of making a system that can track and analyze the action of a human body in real-time and at the same time work on the principle of offline means still images and video.

1.8 Aim of Thesis

The aim of the thesis is to design a model for tracking and activity recognition of human. Working in real-time and offline mode.

1.9 Objective of Thesis

The objective of the thesis is to provide solutions to the problems mentioned, which are the following:

•Using YOLOv5s to solve the problem of tracking systems that do not give an analysis of what the human is doing but only track the human body by building a deep learning system based on the neural network algorithm,

YOLOv5s, which can analyze action and objects and give a detection of human action and track it in two ways, either real-time or offline.

1.10 layout of Thesis

This consists of four other chapters and as follows arranged from these chapters:

- ***Chapter Two:*** *The theoretical underpinning of the general algorithm and approaches employed in this thesis is presented in this chapter.*
- ***Chapter three:*** *This chapter describes the suggested developed system, its associated algorithm in detail, and the actions involved.*
- ***Chapter four:*** *This chapter contains the results acquired after using the suggested system on the data set in question and a commentary on the results.*
- ***Chapter Five:*** *This chapter summarizes the findings of the study and makes recommendations for future research*