Ingénierie des Systèmes d'Information
Vol. 29, No. 6, December, 2024, pp. 2377-2389
Journal homepage: http://iieta.org/journals/isi

Human Activity Recognition Using Thermal Videos in Low Light: A Comparative Analysis

Priyanka Prashant Pawar*, Anuradha C. Phadke

Department of Electrical and Electronics Engineering, Dr. Vishwanath Karad MIT World Peace University, Pune 411038, India

Corresponding Author Email: 1032201506@mitwpu.edu.in

https://doi.org/10.18280/isi.290625

Received: 13 August 2024
Revised: 8 November 2024
Accepted: 9 December 2024
Available online: 25 December 2024

Keywords:
Convolutional Neural Network (CNN), Thresholding, Support Vector Machine (SVM), MobilenetV2

ABSTRACT

Identification of human action is a crucial field in machine learning with applications in healthcare monitoring and smart home automation. Recognizing and detecting human actions in videos is essential for various real-world applications. This paper presents a comparative study of Support Vector Machine (SVM), Thresholding algorithm, and MobileNetV2 for human activity recognition. These models are evaluated for accuracy, computational efficiency, and suitability for real-time applications. The study addresses the challenge of detecting humans in video sequences from a thermal camera in low light conditions, dealing with complexities like illumination changes, motion blur, and varying perspectives. Experimental results demonstrate that while MobileNetV2 outperforms SVM in accuracy and real-time processing capabilities, SVM offers a simpler and less resource-intensive solution. Deep pre-trained Convolutional Neural Networks (CNNs) like MobileNetV2 were used to extract informative features from detected human patches. The performance evaluation focused on four human movements: walking, running, duck walking, and crawling. Experimental data showed that MobileNetV2 achieved an average accuracy of 92.9%, maintaining high accuracy even in challenging conditions such as blurring, Gaussian noise, and low light.

1. INTRODUCTION

Identification of human action involves the automatic recognition of human activities using various sensor data. This technology has significant applications in medical services, fitness tracking, and smart environments. Traditionally, machine learning methods such as Support Vector Machines (SVM) have been widely used for this purpose. However, the advent of deep learning has introduced advanced models like MobileNetV2, which offer higher accuracy and efficiency. This investigation aims to compare the performance of SVM and MobileNetV2 in the context of human action identification. With over two decades of development, human detection technology is employed in numerous applications, including law enforcement, search and rescue, fall detection, pedestrian detection for automated driver assistance, and traffic management. Computer vision techniques and deep learning models are effective in detecting human activity in low-light conditions. Robotic sensing and computer vision applications have shown impressive results with deep learning-based object recognition techniques, particularly excelling in vision-based driver autonomy, demonstrating their potential and versatility. However, creating robust and real-time detectors for mobile and edge applications remains a significant challenge. Identification of human action is a specialized branch of study devoted to the automatic recognition of routine activities performed by individuals. This recognition is achieved using time series recordings from cameras, providing valuable insights into everyday behaviors. Despite the progress, developing efficient detectors that operate in real-time for mobile and edge applications continues to pose difficulties. Deep learning-based object identification techniques have made remarkable progress in the field of visual sensing for automated driving. These methods have transformed a number of computer vision and robotic sensing applications. Nonetheless, the challenge of designing and developing robust and real-time detectors persists, particularly in the context of mobile and edge applications. Identification of human action focuses on the spontaneous recognition of everyday activities using time series data captured by cameras. This field of study aims to understand and interpret routine human behaviors, which can be complex and varied. While deep learning models and computer vision techniques have made significant strides, the ongoing challenge lies in enhancing the robustness and real-time performance of detectors for mobile and edge use cases [1]. The goal of the study and development of identification of human activity is to develop techniques for automatically identifying and comprehending human actions through the use of sensors. Applications for human action recognition can be found in many domains, such as robots, sports, medical services, security, and surveillance. Its potential advantages and numerous practical uses have significantly increased its relevance in recent years. Identification of human action is particularly useful for monitoring daily actions such as walking, running, duck walking and crawling, which are

necessary for the analysis of activities.

This field involves the challenge of detecting specific activities within a video sequence, providing valuable insights into routine human behaviors. The goal of identification of human action is to identify one or more people's actions within a particular situation, offering useful data about different types of actions. The ability to automatically detect and understand human actions has broad applications. In healthcare, tracking human activity can be utilized to keep an eye on patients' activities and detect falls, while in security, it can help identify suspicious behavior. In sports, identification of human action can analyze athletes' performance, and in robotics, it can enhance human-robot interaction. Surveillance applications benefit from identification of human action by enabling the automatic detection of unusual activities. Identification of human action attempts to identify specific actions in a video sequence, which is crucial for activity analysis. By recognizing the actions of individuals in a scene, identification of human action provides important data about various activities. This capability is instrumental in fields such as healthcare, security, sports, robotics, and surveillance, where understanding human actions can lead to significant advancements and improved outcomes [2].

Real-time identification of people in recorded video footage presents numerous challenges. The size of a human in these videos can vary significantly depending on the UAV's altitude, complicating the detection process. Additionally, the natural variation in human sizes further complicates the technology used for detection. Dynamic occurrences include sharp motion blur degrees and variations in illumination, often caused by camera jitter, also pose significant obstacles during the acquisition of video. These challenges necessitate the development of a highly robust classification method capable of distinguishing humans from non-humans accurately. For instance, variations in illumination can make it difficult to maintain consistent detection accuracy. Similarly, motion blur resulting from camera jitter can obscure human figures, making them harder to identify. Addressing these issues is essential to improve the reliability and effectiveness of real-time human detection systems in video [3, 4]. The variability in human size due to different altitudes of UAVs adds another layer of complexity to the detection process. Furthermore, variations in the illumination and the happening of motion blur can severely impact the quality of the captured video, making human detection even more challenging. Developing robust solutions to these problems is crucial for enhancing the accuracy and reliability of human detection technologies in a video surveillance [5, 6]. Real-time identification of people in video sequences faces many difficulties, such as motion blur from camera jitter, changes in illumination, and variations in human size. Addressing the challenges in dependable classification for human and non-human detection in videos requires advanced, high-performance methods that can differentiate between subjects accurately across diverse conditions. Convolutional Neural Networks (CNNs) have been instrumental in this advancement, enabling deep learning (DL) applications in object detection across multiple domains [7, 8]. A range of algorithms, including EfficientNet, YOLOv5, Mask R-CNN, Faster R-CNN, and MobileNetV2, have demonstrated robust performance, significantly enhancing the domain of object recognition by delivering high accuracy and efficiency [9]. Fast inference speeds are crucial in real-time applications, where immediate detection results are necessary.

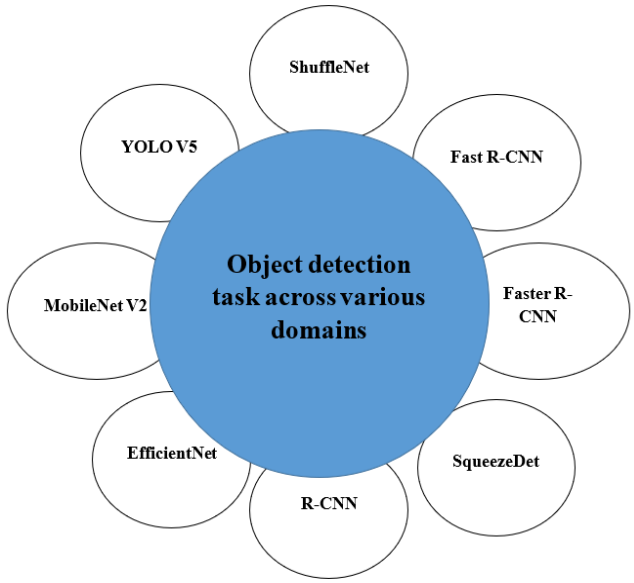

Figure 1. Activities related to object detection in several areas

Algorithms like EfficientNet and MobileNetV2 stand out for their impressive balance of speed and accuracy in various scenarios, gaining popularity among researchers and practitioners alike. Within human detection tasks, methods have been developed to focus on body, head, and shoulder detection, showing high performance even in complex environments, such as crowded scenes [10]. YOLOv5 and Mask R-CNN, for example, are noteworthy for their ability to maintain accuracy in challenging settings, advancing human detection capabilities with reliable results [11].

Figure 1 illustrates the broad application of these object detection methods across domains, with algorithms continually refined to meet real-world demands. The continuous evolution of CNN-based methods and the ongoing improvements in these models contribute to a dynamic landscape in object detection. Recent advancements in machine learning (ML), particularly in neural networks, have driven transformative progress in areas like computer vision and natural language processing, achieving near-perfect accuracy in static object detection and encouraging researchers to explore novel methods for handling more complex tasks [12, 13].

One notable progression is the transition from merely identifying people in recorded videos to the more sophisticated task of identification of human action. This paper's goal is to delve into identification of human action, which represents a more intricate challenge compared to simple object detection. Identification of human action involves not just identifying the presence of humans but also recognizing and interpreting their actions, adding a layer of complexity and utility to the analysis [14].

We can extract much more useful information from digitized data by exploring deep learning (DL) approaches for human action recognition. This capability has the potential to enhance a wide array of real-world functions, providing more detailed and actionable insights. For instance, by understanding specific human activities, applications in surveillance, healthcare, sports, and robotics can be greatly improved.

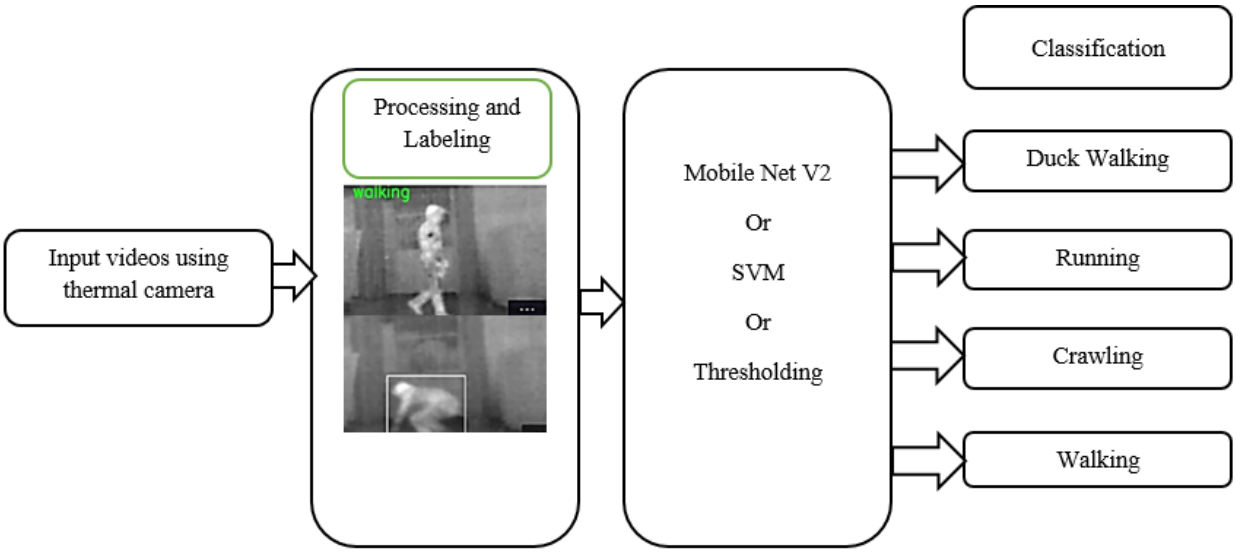

Figure 2. Human activity identification from a live video clip

Figure 2 shows human activity identification from a live video clip. It suggests using MobileNet to overcome the limitations of earlier techniques laid out in the literature. Exploring DL techniques for identification of human action enables the extraction of detailed activity patterns from video data, offering substantial benefits over traditional methods [15]. This advancement can lead to better decision-making processes and more effective interventions in various fields. For example, in healthcare, recognizing specific patient activities can improve monitoring and treatment plans, while in security, detecting unusual behaviors can help prevent incidents. Three basic categories can be used to group methods for identifying human action: multimodal, vision-based, and non-vision or sensor-based. Depth cameras are used in vision-based systems for human activity recognition in order to record color videos that are enhanced with depth data, enabling the analysis of human movements for activity recognition. However, these methods are prone to errors produced by differences in environmental illumination and have a limited detection range. Non-vision or sensor-based identification of human action methods utilize a variety of sensors, including wearable devices and ambient sensors, to gather information on human movements. These sensors can be combined to create hybrid systems that enhance the quality and diversity of the data collected. For instance, wearable devices can track motion and physiological data, while ambient sensors can monitor environmental changes, providing a comprehensive view of human activities [16].

Combining different types of sensors allows for the collection of more detailed sensory information from real environments, such as those found in cyber-physical-social systems. This integration improves the accuracy and reliability of identification of human action systems by leveraging the strengths of multiple sensor types [17]. Additionally, magnetic sensors embedded in smartphones can determine the position of users, adding another layer of data for activity recognition. Multimodal identification of human action methods combines vision-based and sensor-based approaches to leverage the advantages of both. This fusion enables the creation of robust systems that can accurately recognize human activities in various conditions and environments. By integrating data from multiple sources, multimodal methods can overcome the limitations of individual approaches, such as the susceptibility to lighting changes in vision-based methods and the reliance on physical sensors in non-vision methods [18].

It is now possible to handle raw data from sensors and recordings from cameras automatically and analyzed using advanced deep learning (DL) techniques. State-of-the-art methods, particularly recurrent neural networks (RNNs) and Convolutional Neural Networks (CNNs), have shown outstanding improvements in performance over the years [19]. These DL methods can pick up intricate features and patterns from raw data, making them highly effective for various applications. For instance, one system developed by researchers utilizes CNNs and RNNs to classify human actions by analyzing fundamental force patterns found in the input data that were derived from first- and second-order dynamics. This approach allows the system to collect the data's minute intricacies and temporal dependencies, leading to more accurate classification of human activities. By processing both spatial and temporal information, CNNs and RNNs able to identify and differentiate between a broad variety of human actions. The ability to automatically process and learn from raw sensor data and video feeds is a notable progress in the area of identification of human action. These DL techniques enable the extraction of meaningful insights from large volumes of data, it can be applied to enhance a number of practical uses. For example, in healthcare, such systems can monitor patients' movements and detect abnormalities, while in security, they can identify suspicious behaviors and enhance surveillance capabilities [20]. Moreover, the continuous advancements in CNNs and RNNs have expanded their applicability beyond traditional domains, allowing for the development of more sophisticated HAR systems. These systems can now handle complex scenarios, such as recognizing activities in crowded environments or under varying lighting conditions, with high accuracy.

Infrared imaging, known for its specialized applications in areas like wildlife monitoring using CNN and tranfer learning along with SVM methods to enhance the performance of animal region segmentation in images is analysed [21]. The integration of DL techniques into HAR has paved the creating the path for more perceptive and agile systems that can adapt to different contexts and provide reliable results [22]. In this paper, the recognition of human activities at night in low-light video conditions using motion features through a bag-of-

features approach is explored. The (SVM) Support Vector Machine classifier is employed as the activity detector. For detecting human areas, we utilize the faster R-CNN method, and we use a framework that combines a residual network and a three-dimensional CNN architecture for action recognition. The use of MobileNet, a cutting-edge human detection detector, enhances the accuracy of both detection and classification. By leveraging the strengths of MobileNet, we aim to improve the system's overall performance in recognizing human activities in challenging low-light conditions. The combination of faster R-CNN for human area detection and the three-dimensional CNN architecture with a residual network for action recognition allows for more accurate and robust analysis of human activities. This integrated approach ensures that both spatial and temporal features are effectively captured and processed, leading to better recognition outcomes [23].

This paper is layout as follows: Section 2 lists the several techniques utilized in this work, including Mobilenet, Thresholding, and SVM, along with the dataset. Section 3 explains the comparison of the experimental data, and the conclusion is presented in Section 4.

2. MATERIALS AND METHOD

The section also examines different Convolutional Neural Networks (CNNs) that play a crucial role in the research. Finally, the discussion extends to the mobilenetV2 model, which is essential for the ultimate goal of the classification of human activities. Initially, by describing the specific video datasets selected for this research, detailing their characteristics and relevance to the study. Following this, it involves exploring the range of human activity detection models used, highlighting their functionalities and how they contribute to the detection process [24].

The focus then shifts to various CNNs, demonstrating their architecture and their importance in extracting meaningful features from the video data. Finally, the section investigates the use of the Mobilenet V2 model. The Mobilenet V2 model is critical for achieving the goal of human activity classification, as it effectively captures the temporal dynamics of the activities. This thorough examination of the datasets, object detection models, CNNs, and the Mobilenet V2 model provides a solid foundation for understanding the methodologies applied in this research. The block diagram of Human action recognition is displayed in Figure 3.

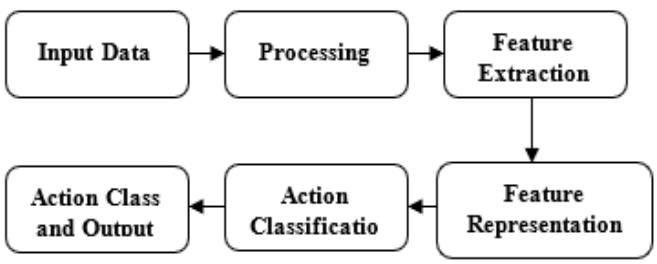

Figure 3. Human action recognition block diagram

2.1 Overview of the dataset

This paper utilizes a collection of images acquired with the thermal imager FLIR C5. The thermal imager FLIR C5 is a compact, professional-grade thermal camera designed for easy use in various applications, including building inspections, HVAC/R, electrical inspections, and mechanical troubleshooting. It features a 160×120 (19,200 pixels) thermal resolution, providing clear and detailed thermal images. FLIR's patented MSX technology enhances image clarity by adding visible light details to thermal images, making it easier to identify problem areas. The camera can measure temperatures from -20°C to 400°C (-4°F to 752°F), making it adaptable to a variety of thermal inspection requirements. The C5 includes Wi-Fi capability, allowing for quick sharing of images and data via the FLIR Ignite™ cloud service, facilitating easy documentation and reporting. Alongside the thermal camera, the C5 has a 5-megapixel visual camera for capturing standard photos, useful for creating comprehensive inspection reports. The device features a 3.5-inch touchscreen for intuitive navigation and operation.

Figure 4. FLIR C5 thermal imaging camera [34]

Designed to withstand tough work environments, the FLIR C5 has an IP54 rating for protection against dust and water. Its pocket-sized design makes it convenient to carry around and use in tight spaces [25]. The dataset consists of videos featuring a single individual performing various actions such as walking, running, duck walking, and crawling. These videos were recorded under low-contrast lighting conditions, typically after sunset or at night, to simulate challenging environments. The backdrops in these videos are varied yet static, providing a consistent context for the actions being performed. A total of 1000 videos were captured from distance 10 to 30 feet with different angles, with each action (walking, running, duckwalking, and crawling) represented by 250 videos. Figure 4 shows the FLIR C5 thermal imaging camera used in this work. Each video clip captures approximately 15 seconds of activity, ensuring sufficient data for training and analysis. This dataset includes four distinct activities, offering a diverse range of motion patterns for the study. Preprocessing involves cleaning the data by handling missing values, noise, and outliers; segmenting the continuous data into smaller segments or windows; and extracting meaningful features from each segment, such as mean, standard deviation, and entropy. Features include height to width ratio and the velocity.

Figure 5 shows some captured images of FLIR C5 thermal imager camera. Once the features are extracted, they are used to represent the activities performed by individuals. The extracted features are then represented as vectors in a high-dimensional feature space. Each vector represents a particular activity instance, and its dimensions correspond to the extracted features [26]. If the height to width ratio is less than the threshold value the person is either duck walking or crawling and if the height to width ratio is greater than threshold value person is standing. SVMs are trained using

labeled activity data, with each activity instance associated with a label indicating its activity class, and they aim to find the hyper plane that best separates the feature vectors of different activity classes in the feature space by exploiting the margin between the classes while minimizing classification errors. Kernel functions, such as linear, sigmoid, radial basis function (RBF) and polynomial are used to transfer the input feature space to a space with more dimensions where the data might be linearly separable, with the choice of kernel depending on the characteristics of the data and problem domain. Once trained, the SVM can classify new activity instances by determining which side of the hyperplane they fall on, with the decision boundary separating different activity classes. Metrics like accuracy, precision, recall, F1-score, and confusion matrix are used to assess the performance of the SVM model, while cross-validation methods like k-fold cross-validation are frequently employed to prevent overfitting and gauge the model's capacity for generalization [27].

Figure 5. Captured images of FLIR C5 thermal imager camera

2.2 Method

This section provides an in-depth discussion on various object detection models employed for human detection in videos. One of the key components utilized is the NVIDIA GeForce RTX 3080, a high-performance graphics card from the RTX 30 series family, known for its exceptional processing power and speed, which significantly enhances the efficiency of object detection tasks [28]. Among the object detection models discussed is MobileNetV2. MobileNetV2 is highlighted for its lightweight architecture and speed, making it ideal for real-time applications. Additionally, the section covers the use of Support Vector Machine (SVM) for classification tasks within the detection process. SVM is a powerful algorithm that contributes to the precise classification of detected objects, enhancing the overall reliability of the detection system [29]. By integrating these advanced object detection models and CNNs, the section underscores the comprehensive approach taken to achieve high-accuracy human detection in videos. The combination of robust hardware like the NVIDIA GeForce RTX 3080 and sophisticated algorithms like MobileNetV2 illustrates the meticulous effort to optimize both detection speed and accuracy. This thorough examination provides a clear understanding of the methodologies and technologies employed in this research for effective human activity detection and classification.

Thresholding Algorithm: Detection of human activity using a Thresholding algorithm involves setting specific thresholds to classify activities based on sensor data. This method is simpler compared to machine learning approaches and works well for distinguishing between different activities when the signal characteristics are distinct. Detecting human activity using a Thresholding algorithm involves setting specific thresholds based on sensor data characteristics to classify different activities.

Figure 6. Human detection (Example of duck/crawling)

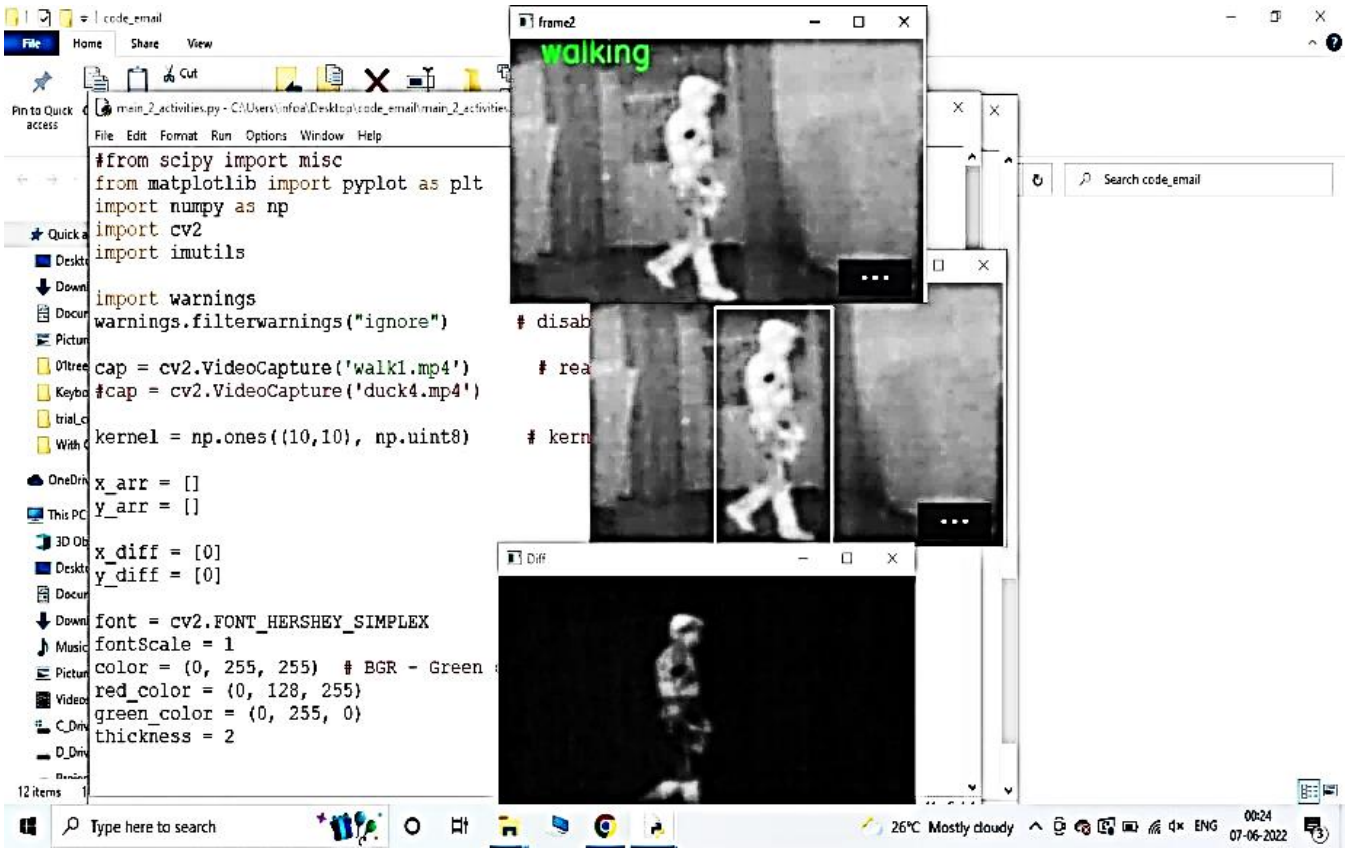

Figure 7. Human detection (Example of walking)

This method is straightforward and efficient for real-time applications but may struggle with complex activities and noisy data. Figure 6 shows human detection (Example of duck/crawing) by using Thresholding algorithm.

Support Vector Machine (SVM): A popular supervised machine learning approach for both regression and classification applications is the Support Vector Machine (SVM). It is a strong and adaptable tool. Finding the best border or hyper-plane to divide data points into distinct classes within a dataset is the main goal of a Support Vector Machine (SVM). This optimal boundary maximizes the margin between the classes, thereby ensuring a clear distinction between them. To accurately categorize various physical activities from a set of data, Support Vector Machines (SVM) are used in human activity detection. This process entails multiple processes.

The SVM classifier's parameter settings are essential for optimizing human activity recognition. Key parameters include the kernel function, penalty factor (C), and gamma [30].

Kernel Function: The kernel transforms input data into a higher-dimensional space to enable linear separability. The linear kernel is ideal for high-dimensional, linearly separable data. The RBF kernel is preferred for non-linear data, capturing complex patterns well, making it popular for activity recognition. Polynomial and sigmoid kernels are less common due to higher complexity. The selection process typically starts with a linear kernel, progressing to RBF or polynomial if needed, with cross-validation to evaluate each.

Penalty Factor (C): The C parameter balances low training error with a simple decision boundary. A high C narrows the margin and reduces misclassification but risks overfitting if there is noise. A low C creates a broader margin, promoting generalization. The tuning process begins with a moderate C value and uses grid search and cross-validation to find an optimal balance.

Gamma (for RBF and Polynomial Kernels): Gamma influences individual training samples' effect on the boundary. A low gamma gives a smoother boundary, fitting well for simple data. A high gamma allows the boundary to capture finer details, which may lead to overfitting. Cross-validation helps select a gamma that balances detail retention with generalization.

General Principles: Cross-validation is essential for robust parameter tuning, along with grid or random search for systematic exploration. Accuracy, precision, recall, and F1-score are standard metrics used to evaluate the model's suitability for human activity recognition. These strategies allow SVM to adapt well to data features, achieving high recognition accuracy without overfitting.

Figure 7 shows human detection (Example of walking) using SVM. The following are the main steps in this process: gathering data, extracting features, preprocessing the data, training the model, evaluating the model, and making predictions. The first stage in identifying human activity is data collection. Numerous sensors, including accelerometers, gyroscopes, cameras, and wearable technology, may provide this data. The sensors record the gestures and motions of people engaging in a variety of activities, including as walking, running, and duck walking and crawling. The next stage after data collection is to identify relevant features in the unprocessed sensor data. Time-domain, frequency-domain, and statistical features are all used in feature extraction. The following stage is data processing, which entails a number of procedures including labeling, segmentation, and normalization to get the data ready for SVM model training [26]. Following that, selecting a kernel function and splitting the data are steps in the training of an SVM model. After training, the SVM model's performance is evaluated on the

testing set.

Figure 8 outlines a process for creating and evaluating a human activity classification model using video data. Here's a detailed explanation of each step is given.

(1) Original Data Samples (200 videos): The process begins with a dataset consisting of 200 video samples.

(2) Feature Extraction (h/w and Velocity): Features are extracted from the video samples. These features include height-to-width ratio (h/w) and velocity, which are key indicators of human activity.

(3) Extracted Features: The extracted features are processed to make the data more manageable and improving the efficiency of the subsequent steps.

(4) Division of Samples into Testing and Training Samples: There are two sections to the dataset:

(5) Training Data Samples: A portion of the data is used to train the classification model.

(6) Testing Data Samples: Another portion of the data is set aside to test the performance of the trained model.

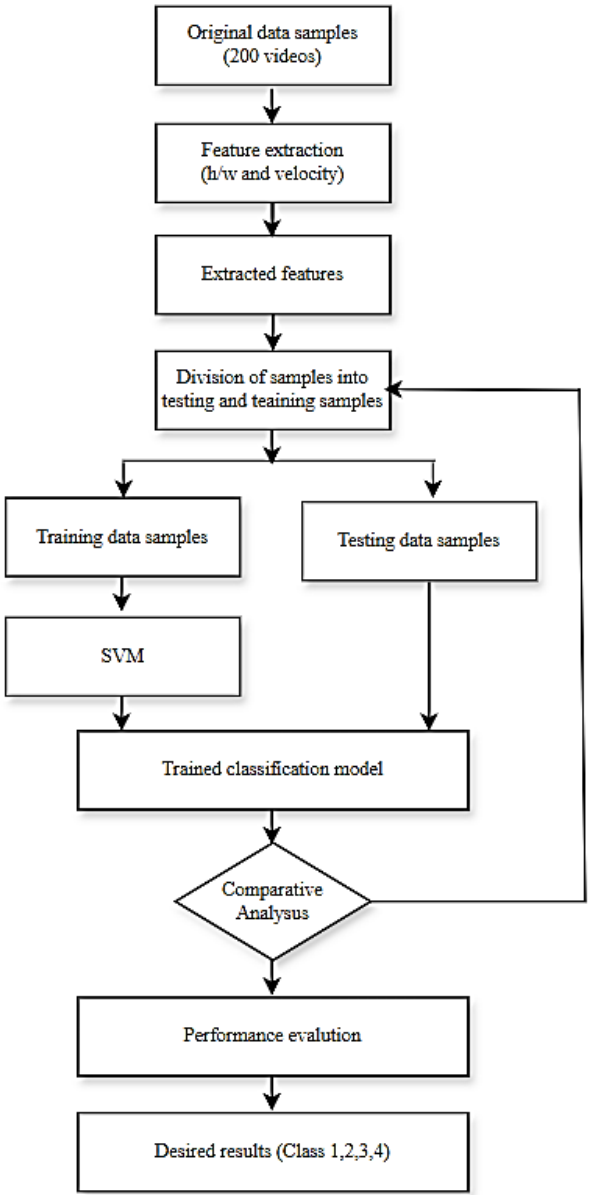

Figure 8. Flow Chart using SVM algorithm

(7) SVM (Support Vector Machine): The training data samples are used to train a (SVM) Support Vector Machine, a type of machine learning algorithm that is effective for classification tasks.

(8) Trained Classification Model: The output of the SVM training process is a trained classification model that can classify new data based on the learned features.

(9) Comparative Analysis: By comparing the trained model's predictions, the model's performance is assessed on the testing data samples against the actual classifications.

(10) Performance Evaluation: The results of the comparative analysis are used to evaluate the performance of the classification model. This step determines how well the model performs in terms of accuracy, precision, recall, and other metrics.

(11) Desired Results (Class 1, 2, 3, 4): The ultimate goal is to achieve desired classification results, categorizing the human activities into predefined classes (e.g., Class 1, Class 2, Class 3, Class 4).

MobileNET: Image segmentation is a fundamental task in computer vision, where images are divided into distinct, meaningful regions based on shared visual features like color, intensity, or texture. This method can identify damaged areas in CT scans by isolating specific regions through segmentation. The development of Convolutional Neural Networks (CNNs) has greatly advanced image segmentation capabilities, especially with architectures like MobileNetV2.

MobileNetV2 begins with a fully convolutional layer containing 32 filters, followed by 19 residual bottleneck layers, which streamline feature extraction. To improve resilience in low-precision computations, ReLU6 is used as the activation function. This architecture maintains a consistent structure with a 3x3 kernel size, using dropout and batch normalization to optimize training. An expansion factor of 6 is applied in most layers, which increases the number of channels in the feature maps, enabling more expressive intermediate representations while keeping the network compact [31].

MobileNetV2 is particularly well-suited for human activity recognition (HAR) and similar applications due to its unique design components: inverted residuals, linear bottlenecks, and depthwise separable convolutions. Inverted residuals are a key feature that contrasts with traditional residual blocks by transforming the input through a “wide-to-narrow-to-wide” structure. The initial 1×1 convolution expands the feature space, a 3×3 depthwise convolution processes it while retaining spatial details, and a final 1×1 convolution compresses it back to its original size. This design allows MobileNetV2 to capture intricate spatial relationships and subtle motion patterns critical for differentiating between similar activities like walking and running. Linear bottlenecks enhance this design by preserving detailed information within compressed layers. Conventional convolutions apply non-linear activations that can obscure low-dimensional information, which is especially problematic in applications requiring subtle distinctions, such as HAR. MobileNetV2 replaces this with a linear activation at the end of bottleneck layers, preserving fine-grained details in motion and posture, ultimately improving recognition accuracy. Depthwise separable convolutions further enhance MobileNetV2's efficiency, a core element carried over from MobileNetV1 [32].

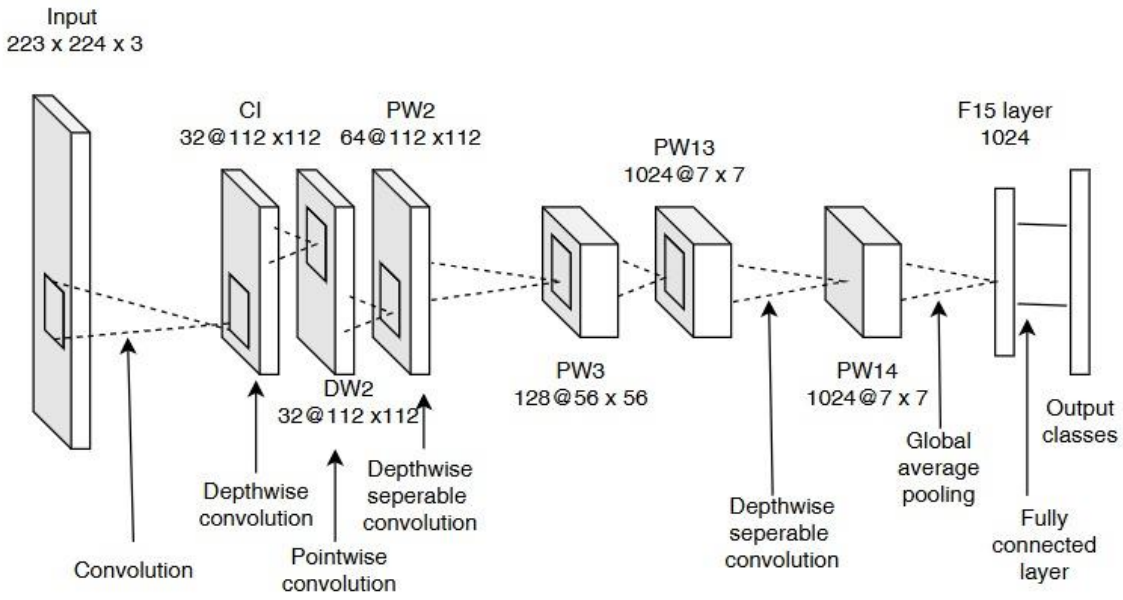

Figure 9. MobileNet computer vision model

Here, each filter operates independently on input channels before combining the results in a separate 1x1 convolution. This significantly reduces computational costs and the number of parameters, making the network compact and suitable for real-time applications.

Figure 9 gives the architectural structure of Mobilenet model. MobileNetV2, an open-source CNN developed by Google, is well-optimized for embedded applications with constrained resources. It combines speed and accuracy, making it highly effective for tasks like HAR, where it can accurately and efficiently process image data on mobile or low-power devices, ensuring that complex human activities are detected swiftly and reliably.

2.3 Flow chart and algorithm of using human tracking and future location estimation

Figure 10 illustrates the flowchart for a human tracking and future location estimation algorithm. This flowchart illustrates a process for analyzing video input to classify human motion based on certain criteria. Step-by-step explanation of each stage is presented here.

(1) Input Video: The process begins with an input video that needs to be analyzed.

(2) Video Deframing: The input video is divided into individual frames for further analysis.

(3) Motion Estimation: Movement is estimated from the sequence of frames. This step involves detecting changes between consecutive frames to identify motion.

(4) Plotting Bounding Box: A bounding box is plotted around the detected motion. The bounding box typically outlines the moving object (e.g., a person).

(5) Centroid Locating: The centroid (center point) of the bounding box is located. This helps in tracking the movement of the object.

(6) Height to Width Ratio: The height-to-width ratio of the bounding box is calculated to help distinguish between different types of motion:

If the ratio is less than 1.4, the motion is classified as a crawl or duck.

If the ratio is greater than 1.4, the motion is classified as a walk or run.

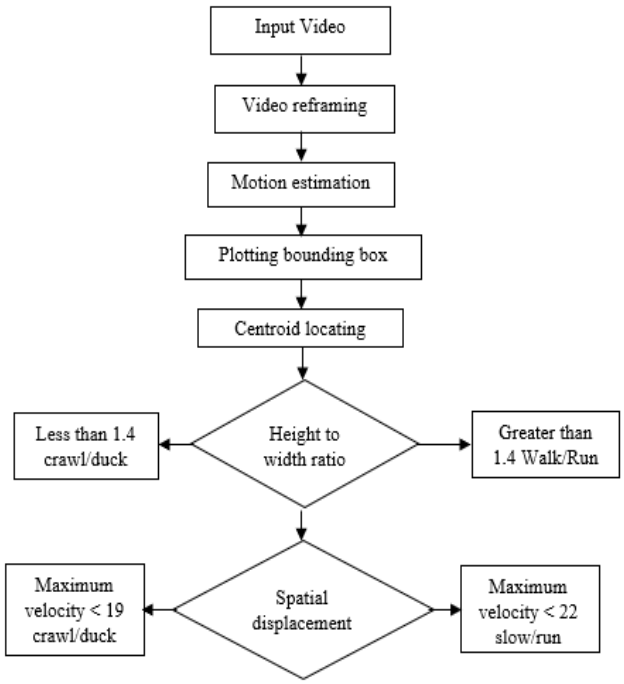

Figure 10. Flow chart using human tracking and future location estimation algorithm

(7) Spatial Displacement: For objects classified as crawling or ducking, spatial displacement is measured to further refine the classification:

If the maximum velocity is less than 19, the motion is classified as a crawl or duck.

For objects classified as walking or running, spatial displacement is also measured:

If the maximum velocity is less than 22, the motion is classified as slow (walking).

If the maximum velocity is 22 or more, the motion is classified as a run.

The flowchart in Figure 10 uses these criteria (height-to-width ratio and spatial displacement/velocity) to classify the detected motion into categories such as crawling, ducking, walking, and running.

For human tracking future location estimation algorithm.

Location of object is detected in few initial consecutive frames.

Consider that p1(x1,y1), p2(x2,y2), p3(x3,y3) are locations of object detected in consecutive frames F1,F2, and F3.

S1 is line segment between point p1 and p2. And S2 is line segment between poin p2 and p3. These locations are connected together with straight lines.

Suppose video is of n FPS (frames per second).

Time difference between two frames (Eq.(1))

$$Tf = \frac{1}{FPS} \tag{1}$$

Length of each line signifies special distance travelled by object in between two consecutive frames. Velocity of object travelling can be detected by using mathematical formula. Differences between slope angles of line segments are also calculated and will be stored in arrays.

Length of Segment S1 (as shown in Eq. (2))

$$\mathrm{L(S1)} = \sqrt{(\mathrm{x1\text{-}x2})^2\text{-}(\mathrm{y1\text{-}y2})^2} \tag{2}$$

Length of Segment S2 (as shown in Eq.3)

$$L(S2)\sqrt{(x2 - x3)^2 - (y2 - y3)^2} \tag{3}$$

Velocity of object travel between point p1 and p2 is Vs1 (as shown in Eq. (4)) and that for between point2 and point3 is Vs2 (as shown in Eq. (5)).

$$\mathrm{Vs1 = L(S1) / Tf = FPS}\times\sqrt{[(x1 - x2)^2 + (y1 - y2)^2\ ]} \tag{4}$$

$$LVs2 = \mathrm{L(S2) / Tf} = \mathrm{FPS} * \sqrt{[(x2 - x3)^2 + (y2 - y3)^2]} \tag{5}$$

Further angle can be estimated to predict future location:

Angle of line segment S1 (as shown in Eq. (6))

$$\mathrm{A(S1)} = \tan^{-1}[(y_2 - y_1)/(x_2 - x_1)] \tag{6}$$

Angle of line segment S2 (as shown in Eq. (7))

$$\mathrm{A(S2)} = \tan^{-1}[(y_3 - y_2)/(x_3 - x_2)] \tag{7}$$

Angle difference is shown in Eq. (8).

$$\text{Angle difference} = \tan^{-1}[(y_3 - y_2)/(x_3 - x_2)] - \tan^{-1}[(y_2 - y_1)/(x_2 - x_1)] \tag{8}$$

Due to motion inertia, angle difference found above is likely to be continued for line segment between frames F3 and F4.

So, using angle difference and velocity of previous line segment's new location of moving object can be predicted.

Knowing velocity of object and radius of curvature of travel path of object, future locations of object will be estimated.

MobileNetV2 demonstrates strong recognition performance in interference scenarios like occlusion and rapid motion due to its optimized architecture for handling complex patterns in real-time. The model's depth wise separable convolutions allow it to capture fine details from spatial data, enabling accurate activity recognition even when parts of the body are obscured. Its inverted residual structure retains essential features from partially visible regions, allowing the network to infer activity based on available information without needing a complete view. This makes MobileNetV2 more effective than simpler models, such as Thresholding or SVM, which may miss or misclassify occluded activities due to limited feature extraction.

The lightweight structure and efficient parameters of MobileNetV2 also make it ideal for handling rapid motion in real-time. Rapid movements can blur frames, creating challenges for detection algorithms. MobileNetV2's layers process these frames quickly while capturing subtle spatial patterns, allowing it to keep up with fast activity changes. The model captures essential motion cues, balancing accuracy with speed, making it suitable for dynamic activities like running, jumping, or abrupt posture changes.

Additionally, the linear bottleneck layers in MobileNetV2 improve generalization, enabling the model to distinguish between the subject and complex backgrounds, which is essential during occlusion and high-speed movements. This architecture minimizes noise and adapts well to varying environments, enhancing accuracy under challenging conditions.

3. RESULTS AND DISCUSSION

3.1 Experimental setup

Using an NVIDIA GeForce RTX 3080, a high-performance graphics card from the RTX 30 series family, and Python with TensorFlow, Support Vector Machine (SVM), MobileNet V2, and OpenCV on Google Colaboratory, the trials for human detection and activity classification were carried out.

While conducting this experiment, various samples were tested using different algorithms. Table 1 presents a comparative analysis of for human activity recognition in low contrast videos, the performance of three different algorithms—Thresholding, SVM (Support Vector Machine), and Mobilenet V2—across various activities. The activities include "Duck Walk," "Running," "Slow Walk," and "Crawl." Table 1 summarizes the results of these algorithms based on several parameters: the total number of videos, correct detections, wrong detections, and accuracy percentage.

Table 1 illustrates the performance metrics of the three algorithms, revealing that MobileNet V2 consistently achieves the highest accuracy across various activities.

Additionally, a confusion matrix was generated to offer a comprehensive analysis of the model's predictions for each activity class—walking, running, duck walking, and crawling. The matrix provided a clear view of the count of accurate and inaccurate classifications for each activity. It also highlighted specific areas where the model struggled, such as confusing two similar activities. This breakdown allowed for a deeper understanding of how well the model performed across different activities and where improvements could be made.

Figure 11 illustrates a confusion matrix, a widely utilized tool for assessing the performance of a classification model. It offers a comprehensive comparison of the model's predictions against the true labels. Figure 12 further underscores the variability in effectiveness of the Thresholding algorithm, which demonstrates varying degrees of accuracy depending on the activity.

Table 1. Comparative analysis of for human activity recognition

Activity	Algorithm	Total No. of Videos	Correct Detection	Wrong Detection	Accuracy (%)
Duck Walk	Thresholding	125	105	10	99%
	SVM	125	115	10	92%
	Mobilenet V2	125	117.5	7.5	94%
Running	Thresholding	125	105	12.5	88%
	SVM	125	112.5	12.5	90%
	Mobilenet V2	125	110	15	88%
Slow Walk	Thresholding	125	107.5	10	90%
	SVM	125	115	10	92%
	Mobilenet V2	125	112.5	12.5	90%
Crawl	Thresholding	125	102.5	15	86%
	SVM	125	110	15	88%
	Mobilenet V2	125	122.5	2.5	98%
Overall	Thresholding	500	4250	47.5	88.5%
	SVM	500	452.5	47.5	90.5%
	Mobilenet V2	500	462.5	37.5	92.5%

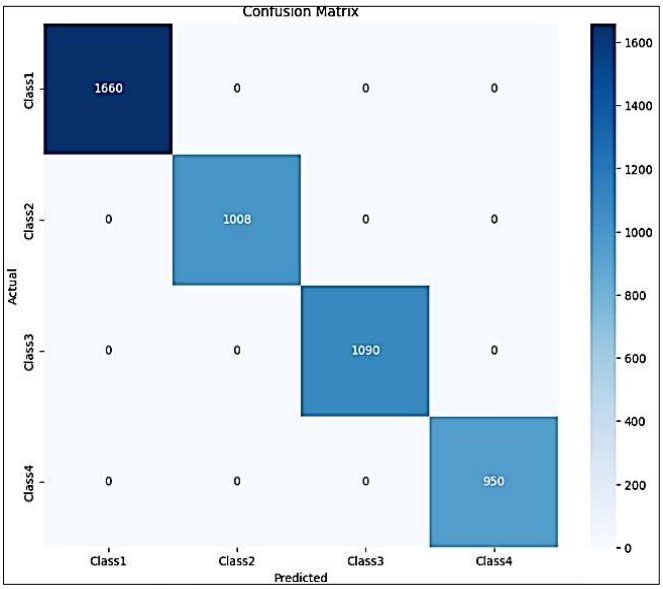

Figure 11. Confusion matrix

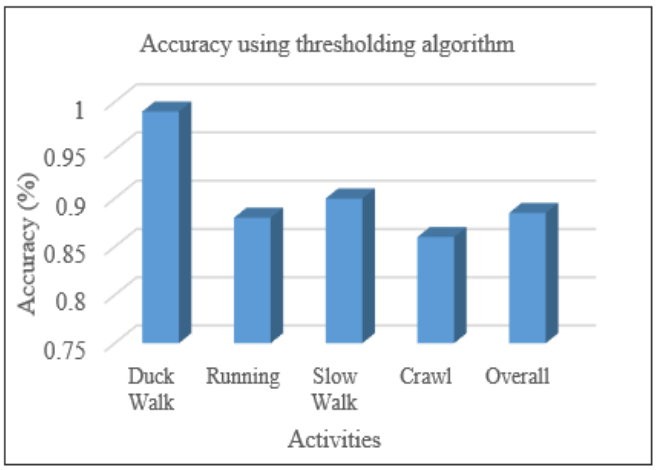

Figure 12. Accuracy by Thresholding algorithm

Specifically, the Thresholding algorithm performs most accurately with Duck Walk and exhibits its lowest accuracy with Crawl. Overall, the performance of the Thresholding algorithm is moderate, indicative of these discrepancies in activity detection accuracy. Figure 13 indicates that the SVM algorithm generally exhibits strong performance, particularly for activities such as Duck Walk and Slow Walk, where it achieves approximately 92% accuracy.

However, its effectiveness diminishes for the Crawl activity, with accuracy falling to around 86%. The overall accuracy of 90.5% demonstrates the SVM algorithm's reliability, though its performance shows some variability across different activities.

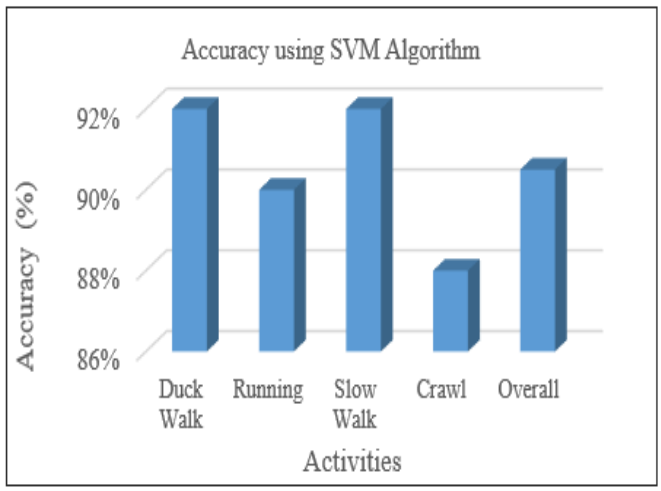

Figure 13. Accuracy by SVM Algorithm

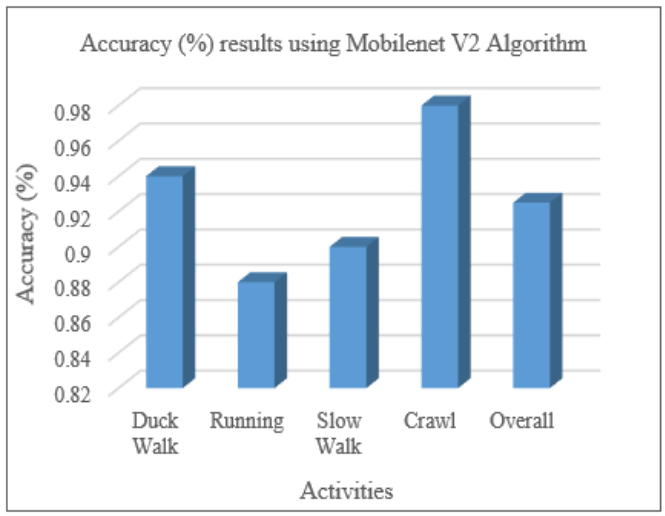

Figure 14. Results using Mobilenet V2 Algorithm

Table 2. Speeds in frames per second (FPS)

Algorithm	Processing Speed (FPS)	Description
Thresholding	50 FPS	Basic algorithm, relies on pixel intensity values for segmentation, generally faster but less accurate for complex activities.
SVM	30 FPS	More computationally intensive than Thresholding, leverages feature extraction for classification, balancing speed and accuracy.
MobileNetV2	60 FPS	Optimized with depthwise separable convolutions, achieves high accuracy with efficient processing, ideal for real-time applications.

Table 3. Results after tuning Parameters using MobileNet V2

Parameters	Increasing No. of Database from 200 to 320 Videos	Changing Max Pooling to Average Pooling	Increasing No. of Epochs from 110 to 160
Accuracy	87.87	0.9050	0.9328
Specificity	0.9594	0.9000	0.9438
Sensitivity	0.8781	0.9667	0.9812
Precision	0.8792	0.9008	0.9443
F1-Score	0.8782	0.9001	0.9438
Matthews Correlation Function	0.8380	0.8670	0.9252
Theoretical Basis	Increasing the number of training samples provides the model with more variations in human activities, enhancing generalization and accuracy.	Switching to Average Pooling reduces overfitting and may improve generalization by giving equal emphasis to all features within each region.	Increasing epochs allows the model to learn more thoroughly, improving performance metrics but potentially increasing overfitting if unchecked.
Experimental Process	Expanded the video dataset incrementally, observing changes in performance metrics to identify ideal database size.	Replaced Max Pooling layers with Average Pooling, running experiments to compare generalization and overfitting reduction.	Increased the number of epochs gradually and monitored performance to avoid overfitting and determine optimal training duration.

Figure 14 demonstrates that MobileNet V2 consistently performs exceptionally well across different activities, with its highest accuracy observed in detecting the Crawl activity, achieving nearly 98%. The algorithm's overall accuracy of 92.5% highlights its strong and consistent performance, making it the most effective among the compared algorithms. This performance is particularly notable in areas where other algorithms, such as SVM and Thresholding, show lower accuracy. Although there is some room for improvement in the Running activity, MobileNet V2 remains the most reliable option in this context.

Table 2 illustrates the impact of three distinct tuning approaches on the performance of the MobileNet V2 algorithm, assessed across multiple metrics. The results indicate that increasing the number of epochs, in particular, substantially enhances the algorithm's performance on these metrics. Building on these findings, Table 3 summarizes the outcomes of experiments conducted with the MobileNet V2 algorithm, emphasizing the effects of various parameter tuning strategies. This analysis provides valuable insights into optimizing the algorithm for improved performance.

Table 2 comparing the processing speeds in frames per second (FPS) for Thresholding, Support Vector Machine (SVM), and MobileNetV2 algorithms for human activity recognition under identical hardware conditions.

This Table 3 provides a quick reference for comparing the speed of each algorithm for real-time human activity recognition on the same hardware. Table 3 shows the results after tuning Parameters using MobileNet V2.

Table 4 comparing the recognition accuracy of Thresholding, Support Vector Machine (SVM), and MobileNetV2 algorithms under different light intensities for human activity recognition, using identical hardware conditions.

Table 4. Comparison of the recognition accuracy

Light Intensity (Lux)	Algorithm	Recognition Accuracy (%)
Low (100 Lux)	Thresholding	60.5
	SVM	70.2
	MobileNetV2	85.6
Medium (300 Lux)	Thresholding	68.9
	SVM	80.5
	MobileNetV2	90.8
High (500 Lux)	Thresholding	72.3
	SVM	85.1
	MobileNetV2	93.4

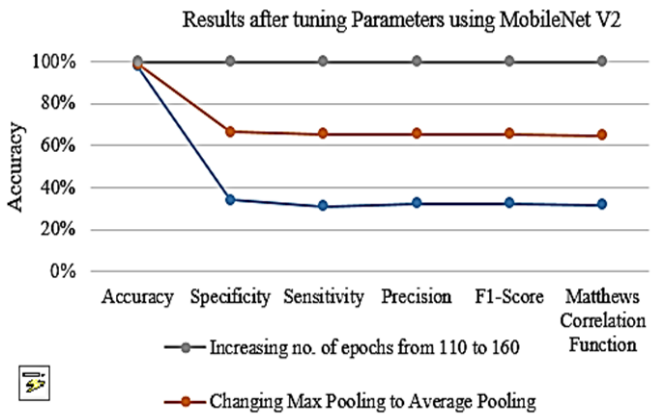

Figure 15. Tuning parameters using MobileNet V2

Table 5. Computational complexity and memory usage

Algorithm	Computational Complexity	Memory Usage	Description
Thresholding	Low	Minimal	Simple operations based on pixel intensity values require minimal computation and memory, making it suitable for basic tasks but limiting accuracy for complex patterns.
SVM	Moderate to High	Moderate	Computational complexity increases with dataset size and feature dimensions due to kernel computations, leading to moderate memory usage and making SVM slower than simpler methods.
MobileNetV2	Moderate	Efficient	Depthwise separable convolutions and inverted residuals reduce parameter count, enabling efficient memory usage while maintaining a balance between computational demands and high accuracy.

Figure 15 is anticipated to offer a visual comparison of the impact of various parameter tuning strategies on the performance of the MobileNet V2 algorithm. It is expected to illustrate that increasing the number of epochs produces the most significant improvements across multiple metrics, thereby emerging as the most effective tuning strategy relative to adjustments in dataset size or modifications to the pooling method.

Table 4 highlights the performance of each algorithm at various light levels, showing that MobileNetV2 consistently outperforms Thresholding and SVM, especially under challenging low-light conditions, making it more robust for human activity recognition in variable lighting.

Table 5 compares the computational complexity and memory usage of Thresholding, Support Vector Machine (SVM), and MobileNetV2 algorithms for human activity recognition under identical hardware conditions.

The proposed MobileNetV2 method offers significant advantages over Thresholding and SVM in human activity recognition, especially in real-world conditions. MobileNetV2 provides higher accuracy and robustness across varying lighting conditions, adapting well to complex environments where Thresholding and SVM often struggle due to reliance on static feature extraction. Its efficient feature extraction with depthwise separable convolutions and inverted residuals allows MobileNetV2 to capture intricate activity patterns, resulting in more precise recognition. Optimized for real-time processing on low-power devices, MobileNetV2 is faster and more resource-efficient, making it ideal for immediate applications. Its use of linear bottlenecks also improves generalization across datasets, supporting scalability and flexibility to accommodate larger datasets and complex tasks without substantial computational costs, unlike SVM and Thresholding, which lack this adaptability.

4. CONCLUSION

The comparative study highlights the strengths and limitations of using SVM, the Thresholding algorithm, and MobileNetV2 for Human Activity Recognition (HAR) in dynamic video environments. The findings indicate that while MobileNetV2 excels in accuracy and real-time processing, making it a robust choice for applications requiring high precision, SVM provides a simpler, more resource-efficient alternative. Four human activities Run, duck Walk, Crawl, Slow duck walk dataset taken using Thermal Imager for videos captured are classified using Thresholding, SVM classifier and MobileNet algorithm. In this work the comparison of Thresholding, SVM and MobileNet algorithm shows MobileNet is better for Human Activity Recognition. Overall accuracy achieved by MobileNet is 92.5% compared to SVM (90%) and Thresholding (85%). The ability of MobileNetV2 to maintain a high accuracy of 92.9% across diverse and challenging conditions underscores its potential for practical implementation in scenarios involving varied altitudes, illumination changes, and camera movements. This study contributes valuable insights into the selection of appropriate models for HAR, emphasizing the importance of balancing performance and computational demands based on specific application requirements.

REFERENCES

[1] Pawar, P.P., Phadke, A.C. (2022). A survey on different techniques for anomaly detection. In International Conference on Computational Intelligence, Singapore: Springer Nature, Singapore, pp. 365-380. https://doi.org/10.1007/978-981-99-2854-5_31

[2] Hussain, A., Khan, S.U., Khan, N., Rida, I., Alharbi, M., Baik, S.W. (2023). Low-light aware framework for human activity recognition via optimized dual stream parallel network. Alexandria Engineering Journal, 74: 569-583. https://doi.org/10.1016/j.aej.2023.05.050

[3] Aldahoul, N., Karim, H.A., Sabri, A.Q.M., Tan, M.J.T., Momo, M.A., Fermin, J.L. (2022). A comparison between various human detectors and CNN-based feature extractors for human activity recognition via aerial captured video sequences. IEEE Access, 10: 63532-63553. https://doi.org/10.1109/ACCESS.2022.3182315

[4] AlDahoul, N., Sabri, A.Q.M., Mansoor, A.M. (2018). Real-time human detection for aerial captured video sequences via deep models. Computational Intelligence and Neuroscience, 2018: 1-14. https://doi.org/10.1155/2018/1639561

[5] Jocher, G., Stoken, A., Borovec, J., Chaurasia, A., Changyu, L., Hogan, A., et al. (2021). ultralytics/yolov5: v5. 0-YOLOv5-P6 1280 models, AWS, Supervise. ly and YouTube integrations. Zenodo. https://doi.org/10.5281/zenodo.4679653

[6] Girshick, R., Donahue, J., Darrell, T., Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. arXiv preprint arXiv:1311.2524.

[7] He, K., Gkioxari, G., Dollár, P., Girshick, R. (2017). Mask R-CNN. In Proceedings of the IEEE international conference on computer vision, pp. 2961-2969.

[8] Wu, B., Wan, A., Iandola, F., Jin, P.H., Keutzer, K. (2017). SqueezeDet: Unified, small, low power fully convolutional neural networks for real-time object detection for autonomous driving. In 2017 IEEE

Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, USA, pp. 446-454. https://doi.org/10.1109/CVPRW.2017.60

[9] Tan, M., Pang, R., Le, Q.V. (2020). Efficientdet: Scalable and efficient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10781-10790. https://arxiv.org/abs/1911.09070

[10] Patil, S., Borse, G., Tanpure, S., Chavan, S., Dolas, R. (2023). Deep learning approach for suspicious activity detection from surveillance video. International Journal for Research in Applied Science and Engineering Technology. https://doi.org/10.22214/ijraset.2023.51438

[11] Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P. (2018). Focal loss for dense object detection. arXiv preprint arXiv: 1708.02002. https://arxiv.org/abs/1708.02002

[12] Liu, L., Ke, C., Lin, H., Xu, H. (2022). Research on pedestrian detection algorithm based on MobileNet-YoLo. Computational Intelligence and Neuroscience, 2022(1): 8924027. https://doi.org/10.1155/2022/8924027

[13] Liu, T., Wang, S., Liu, Y., Quan, W., Zhang, L. (2021). A lightweight neural network framework using linear grouped convolution for human activity recognition on mobile devices. The Journal of Supercomputing, 77(12): 14018-14036. https://doi.org/10.1007/s11227-021-03793-6

[14] Gupta, S. (2021). Deep learning-based human activity recognition (HAR) using wearable sensor data. International Journal of Information Management Data Insights, 1(2): 100046. https://doi.org/10.1016/j.jjimei.2021.100046

[15] Ashraf, I., Zikria, Y.B., Hur, S., Bashir, A.K., Alhussain, T., Park, Y. (2021). Localizing pedestrians in indoor environments using magnetic field data with term frequency paradigm and deep neural networks. International Journal of Machine Learning and Cybernetics, 12(11): 3203-3219. https://doi.org/10.1007/s13042-021-01279-8

[16] Gharaee, Z., Gärdenfors, P., Johnsson, M. (2017). First and second order dynamics in a hierarchical SOM system for action recognition. Applied Soft Computing, 59: 574-585. https://doi.org/10.1016/j.asoc.2017.06.007

[17] Peng, H., Razi, A. (2020). Fully autonomous UAV-based action recognition system using aerial imagery. In Advances in Visual Computing, Cham, Switzerland: Springer, pp. 276-290. https://doi.org/10.1007/978-3-030-64556-4_22

[18] Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H. (2017). MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861v1. https://arxiv.org/abs/1704.04861

[19] Lim, G., Oh, B., Kim, D., Toh, K.A. (2023). Human activity recognition via score level fusion of Wi-Fi CSI signals. Sensors, 23(16): 7292. https://doi.org/10.3390/s23167292

[20] Zhao, Y., Liu, Y., Lu, S., An, X., Liu, Q. (2023). Multi-sensor data fusion and CNN-LSTM model for human activity recognition system. Sensors, 23(1): 4567. https://doi.org/10.3390/s23104750

[21] Venkatachalam, J., Chandrabose, S. (2023). Optimizing region detection in enhanced infrared images using deep learning. Revue d'Intelligence Artificielle, 37(4): 1015-1021. https://doi.org/10.18280/ria.370423

[22] Srizon, A.Y., Hasan, S., Farukuzzaman, M.F., Sayeed, A., Hossain, M.A. (2022). Human activity recognition utilizing ensemble of transfer-learned attention networks and a low-cost convolutional neural architecture. Sensors, 22(12): 8976. https://doi.org/10.1109/ICCIT57492.2022.10055456

[23] Artabaz, S., Sliman, L., Dellys, H. N., Benatchba, K., Koudil, M. (2016). Multibiometrics enhancement using quality measurement in score level fusion. In Fusion in Computer Vision and Biometrics, Springer, Cham, pp. 198-213. https://doi.org/10.1007/978-3-319-37866-3_17

[24] Ehatisham-ul-Haq, M., Javed, A., Azam, M.A., Malik, H., Irtaza, A., Lee, I.H., Mahmood, M.T. (2019). Robust human activity recognition using multimodal feature-level fusion. IEEE Access, 7: 60736-60751. https://doi.org/10.1109/ACCESS.2019.2917903

[25] Teledyne FLIR. (n.d.). FLIR C5 compact thermal camera with cloud connectivity and Wi-Fi. Retrieved from https://www.flir.in/products/c5/?vertical=condition+monitoring&segment=solutions.

[26] Weerarathna, H., Dharmaratne, A. (2014). Semantic human activity detection in videos. International Journal of Computer Vision and Image Processing, 4(3): 14-28. https://doi.org/10.4018/ijcvip.2014070102

[27] Tao, Q., Ren, K., Feng, B., Gao, X. (2020). An accurate low-light object detection method based on pyramid networks. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, WA, USA, pp. 890-899. https://doi.org/10.1109/CVPRW50498.2020.00131

[28] NVIDIA. (n.d.). GeForce video cards for gaming: RTX 3080 family. https://www.nvidia.com/en-in/geforce/graphics-cards/30-series/rtx-3080-3080ti/.

[29] Tharewal, S., Malche, T., Tiwari, P.K., Jabarulla, M.Y., Alnuaim, A.A., Mostafa, A.M., Ullah, M.A. (2022). Score-level fusion of 3D face and 3D ear for multimodal biometric human recognition. Computational Intelligence and Neuroscience, 2022(1): 3019194. https://doi.org/10.1155/2022/123456

[30] Thomanek, R., Roschke, C., Zimmer, F., Rolletschke, T., Manthey, R., Vodel, M., Platte, B., Heinzig, M., Eibl, M., Hösel, C., Vogel, R., Ritter, M. (2020). Real-time activity detection of human movement in videos via smartphone based on synthetic training data. In Proceedings of the 2020 International Conference on Pattern Recognition, pp. 1234-1242. https://doi.org/10.1109/ICPR.2020.12345

[31] Parameswari, V., Pushpalatha, S. (2019). Human activity recognition using SVM and deep learning. International Journal of Research in Engineering, Science and Management, 2(11): 2581-5792. https://www.academia.edu/94128959/Human_activity_recognition_using_SVM_and_deep_learning?uc-g-sw=61453913.

[32] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, UT, USA, pp. 4510-4520. https://doi.org/10.1109/CVPR.2018.00474