[1]P. Naaraju

[2]Dr. Manchala Sadanandam

Advancements in Motion Detection Within Video Streams Through the Integration of Optical Flow Estimation and 3D-Convolutional Neural Network Architectures

Abstract: - Motion detection in video streams is important for many applications, such as autonomous navigation, human-computer interaction, and spying. Conventional techniques, which depend on backdrop removal or frame differencing, frequently break down when dealing with intricate motion patterns and occlusions. Current techniques struggle to handle occlusions and correctly capture complex motion patterns. Furthermore, these techniques could be computationally expensive, especially when dealing with big amounts of video data. An integrated strategy combining optical flow estimation with 3D convolutional neural network (3D-CNN) architectures is presented to address these limitations and boost motion detection systems' accuracy and efficiency. The suggested technique is unusual as it combines 3D CNNs with optical flow estimates. Motion vectors representing dynamic scene changes are produced by optical flow estimates, and spatiotemporal characteristics are extracted by 3D CNNs processing together with optical flow information. By using the complementing capabilities of both approaches, the method performs better in motion detection applications such as object tracking, action recognition, and anomaly detection in video streams. The efficiency of the strategy is assessed by experiments carried out on benchmark datasets. The results show that it is more accurate, resistant against occlusions, and computationally efficient than existing approaches. The suggested approach offers a viable way to improve motion detection capabilities for a range of practical uses. The results show a 98% improvement in processing efficiency and motion detection accuracy when compared to baseline methods. More research and advancement in this area are made possible by the attainable way that MATLAB's suggested approach offers to enhance motion detection skills in a variety of real-world applications.

Keywords: Motion detection, Video streams, Optical flow estimation, 3D-Convolutional neural networks, Computer vision

1. Introduction

The identification of occurrences that vary from anticipated behaviour is known as anomaly detection in videos, and it is a critical task in both video analytics and video surveillance [1]. However, for the following reasons, video anomaly identification is a very difficult task: First of all, genuine video data is intricate, and certain anomalous data points may be located on the edge of regular regions. For example, although skateboarders and walkers have similar appearances, skateboarders are aberrant items that are not allowed on pedestrian walkways. Secondly, there is a shortage of tagged training information for anomaly detection [2]. While anomalous samples are rare and expensive to get, normal patterns are typically rather simple to gather. As a result, anomaly detection techniques may just use the normal data for training their models to identify regularities in an unsupervised environment and identify the instances that vary from the typical patterns [3].

Human Action Recognition (HAR) has many different applications. Its objective is to recognise an individual's actions using either visual or sensor data. There are three types of HAR methods: multi-modal, nonvisual sensor-based, and visual sensor-based [4]. The shape of the felt data is the primary distinction between the visual and various other categories. Some systems record the visual data as 1D signals, while others record the data as 2D, 3D, or video images. Wearable technology has advanced over the past several years, with the development of smartwatches, fitness bracelets, and smartphones [5]. These are outfitted with tiny, non-visual sensors as well as communication and processing power. Additionally, their very inexpensive cost has enabled them to provide new opportunities due to their widespread usage. These consist of illness prevention, rehabilitation training, and health monitoring. Simultaneously, among the most common and hot topics in computer vision studies is visual sensor-based approaches for human action identification [6].

[1]Research Scholar, Department of Computer Science & Engineering, Kakatiya University, Warangal, Telangana, India

Email: nagarajupampati1234@gmail.com

[2]Professor, Department of Computer Science & Engineering, Kakatiya University, Warangal, Telangana, India.

Email: msadanandam@kakatiya.ac.in

Applications include content-based video search, smart video surveillance, environmental assisted living, interaction between humans and robots, and human-computer interaction. The recognizing system is taught to discern between activities performed in a scenario in each of those apps. Based on that inference, it could also make certain judgments or carry out additional processing [7]. It may be said that wearable technology has several drawbacks, including the requirement to be worn and used continuously in the majority of situations. This might provide a serious problem for real-world applications that need to be deployable and ready. Consequently, imposing particular technological specifications about the sensor's performance, size, and battery life, among other things [8]. Furthermore, they might not be effective or appropriate to use in situations such as crowd applications or other similar ones. These restrictions do not apply to HAR which is computer vision-based. The majority of application cases may use computer vision-based HAR without these technological constraints.

Motion is frequently an essential indicator for identifying actions. For instance, since how an action is interpreted is dependent on its direction of motion, it might be challenging to distinguish between two actions, such as "open a door" and "close a door," from a single frame [9]. To address this, current research approaches motion recognition as a separate problem, with one network, the "temporal stream," seeing just a manually created motion representations as input, and another, the "spatial stream," observing the unprocessed RGB video frames. On the other hand, spatiotemporal filters in a 3D Convolutional Neural Network allow the spatial stream to react to motion in the video [10]. Theoretically, as supported by the research, this ought to enable the spatial streams to pick up motion characteristics. Even at this point using a "temporal" 3D CNN which receives as input a clear depiction of motion, and usually optical flow allows us to achieve significant accuracy increases [11]. For example, researchers notice a 6.6% gain in accuracy on HMDB-51 when they ensemble a 3D CNN that captures RGB images with a 3D CNN that collects optical flow frames.

Motion detection in video streams is an essential job with many applications, including autonomous navigation, human-computer interaction, and spying. Accurate real-time tracking and detection of moving objects is essential for maintaining environment security, permitting safe navigation for self-driving cars, and promoting immersive experiences in interfaces between humans and computers. That said, conventional motion detection techniques frequently lack computing efficiency, robustness, and accuracy. Opportunities to get around these restrictions and improve motion detection skills have been made possible by recent developments in deep learning and computer vision. Particularly, 3D-CNNs have demonstrated exceptional performance in extracting discriminative characteristics from visual input, leading to notable gains in a range of computer vision applications. Furthermore, useful temporal information is provided by optical flow estimation, which determines how objects move between successive frames in a video stream. This information may be used to supplement the spatial properties that 3D-CNN structures study.

In this study, a unique motion detection technology that combines 3D-CNN architectures with optical flow estimates. The goal in merging these two approaches is to solve the drawbacks of conventional approaches and make use of their complimentary advantages. 3D-CNNs are excellent at collecting spatial patterns and features, whereas optical flow estimation provides information about the motion dynamics of objects across time. The smooth amalgamation of these methodologies facilitates the identification of moving objects in video streams with greater precision and resilience. The method is new since it uses both 3D-CNN and optical flow estimates in a single, integrated framework to detect motion comprehensively. This integration makes it possible to fully comprehend motion dynamics and makes it easier to make decisions while identifying moving things. To illustrate the efficacy of this technique, provide a comprehensive methodology that includes data collection, pre-processing, model construction, training, assessment, and optimization. Evaluating the performance of the proposed approach through extensive tests on standard datasets, demonstrating notable gains in computing efficiency and motion detection accuracy over baseline approaches. All things considered, this method provides a viable means of improving motion detection skills in a range of practical applications, setting the stage for more study and advancement in this area.

These are some major contributions that have been made:

1. A unique approach that smoothly combines 3D-CNN architectures for motion detection in video streams with optical flow estimation.

2. Using the mutual advantages of deep learning and optical flow estimates to achieve notable gains in motion detection resilience and accuracy.
3. Real-time motion detection solutions are made possible by effectively using 3D-CNN architectures and optimization approaches to address the computational limitations of conventional methods.
4. Offering a comprehensive approach to motion detection by presenting a thorough technique that includes data collection, pre-processing, model development, training, assessment, and optimization.
5. Performing comprehensive tests on reference datasets to verify the efficacy of the suggested methodology, exhibiting its advantages over baseline techniques concerning precision and computing efficiency.

The rest of the research is arranged as follows: The first section presents the subject of motion detection in video streams and explains the rationale for the suggested approach. In Section 2, relevant works in motion detection are reviewed, with an emphasis on the advantages and disadvantages of current methods. In Section 3, the shortcomings of conventional motion detection techniques are discussed, highlighting the need for new approaches. The suggested technique is presented in Section 4, which also describes how 3D-CNN architectures for motion detection are integrated with optical flow estimates. Extensive experiments on benchmark datasets are included in Section 5, together with a description of the experimental setting, data collection, training models, and assessment techniques. Section 6 provides a final summary of the study, a discussion of the implications, and an overview of future research directions for enhancing motion detection skills inside video streams.

2. Related Works

Zhao et al.[12] explains An investigation is conducted on an ongoing Chinese language sign recognition system. The deaf and mute can utilize Chinese sign language, which this technology can recognize. It can output the findings in real time as text. Initially, a Chinese sign language dataset with 500,000 video samples is created using a standard RGB camera. To improve the identification accuracy of the framework for real-time applications, a three-dimensional (3D-CNN) is investigated in combination with optical flow processing using overall variation regularization and L1-norm robust (TV-L1). An equal amount of keyframes from every video stream is extracted using a two-stage down-frame processing method, which is then fed into a 3D CNN to generate feature vectors. On a dataset with 1,000 vocabulary words, comparative studies are done between the recurrent neural networks (RNN) and hidden Markov model (HMM), which achieves 92.6% recognition accuracy. Finally, a fully functional real-time system for recognising sign language is constructed and published. It consists of a video capture mechanism, a motion detection section, a hand and head detection module, and an interface for human interaction. The system's real-time generalization performance is confirmed by the experimental findings. To enhance motion detection, adding skin and human skeleton detection may result in more computing complexity and privacy issues.

The field of human-computer interaction has several applications for which hand gestures are a helpful tool. The goal is to follow the hand's movement in this instance, regardless of the hand's size, shape, or color. A motion template influenced by optical flow (OFMT) is also suggested as a solution for this. OFMT is a condensed depiction of a gesture's motion data contained in a single picture. Comparing deep networks to traditional featurebased approaches that are manually created, significant advancements have been demonstrated recently. Furthermore, it is shown that utilizing several streams with enlightening input data contributes to improving recognition performance. Sarma et al.[13] proposes a hand gesture recognition model using two streams of fusion. The two-stream networks is composed of two layers: a 2D-CNN that receives OFMT pictures as input, and a 3D convolutional neural network (C3D) that receives gesture movies as input. While OFMT helps to remove unnecessary movements and provide more motion information, C3D has demonstrated its effectiveness in collecting the spatiotemporal data contained in a video. To improve the recognition results, every stream is joined utilizing a fusion system, even though each stream may operate separately. Researchers have demonstrated the suggested two-stream network's effectiveness on two datasets.

It is difficult to identify aggressive activity in recordings to maintain security and safety for the general audience. Accurately recognizing and classifying violent episodes in real-world closed-circuit television, which differ in terms of features and locations, requires a thorough comprehension and processing of the sequential data incorporated into these movies. The goal of this work is to present a model that can effectively understand the spatiotemporal setting of films in a variety of environments and violent scenario requirements. Park et al.[14] propose a technique that uses optical flow and RGB data to precisely capture spatiotemporal characteristics connected to aggressive actions. The core network of the method is a Conv3D-based ResNet-3D model, which can process high-dimensional video input. By including an attention mechanism that gives more weight to the most important frames in the RGB and optical-flow sequencing during violent episodes, violence detection becomes more accurate and efficient. The UBI-Fight, Hockey, Crowd, and Movie-Fights datasets were used to assess the model. The results showed that the suggested approach beat current state-of-the-art methods, with area under the curve scores of 95.4, 98.1, 94.5, and 100.0 on the corresponding datasets. Furthermore, this work promises to further a wider range of studies in video analysis and comprehension in addition to having applications in real-time surveillance systems.

Sarma et al.[15] explains In the field of human-computer interaction, hand gestures may be a helpful tool for many different applications. Hand gesture approaches can be especially used in surgical robotics, virtual and augmented realities, sign language recognition, and many other fields. Because hands vary in size and form, it might be difficult to identify and track moving hands accurately throughout the hand gesture identification process. Here, tracking the hand's motion is the main goal, regardless of the hand's size, shape, or colour. A motion template driven by optical flow is also suggested as a solution for this. OFMT is a condensed depiction of a gesture's motion data contained in a single picture. Several datasets were utilized in the experiment, one with a naked hand with an open palm and the other with a folded palm while wearing green gloves. In both scenarios, they were able to produce OFMT pictures with the same level of accuracy. Comparing deep network-based methods to traditional feature-based methods that are manually created, significant advancements have been observed recently. Furthermore, it is observed in the literature that using several streams with useful input data improves recognition accuracy performance. This paper essentially suggests a simple yet effective motion template that utilises optical flow as well as a two-stream fusing model for hand gesture identification. The twostream network is composed of two layers: a 2D-CNN that receives OFMT pictures as input and a 3D convolutional neural network (C3D) that receives gesture movies as input. C3D is effective in capturing the spatiotemporal details of a video. On the other hand, OFMT offers more motion information while assisting in the removal of unnecessary motions. To improve the recognition results, each stream is joined utilizing a fusion system, even though each stream may operate separately. Researchers have demonstrated the suggested twostream network's effectiveness on two datasets.

Peng et al.[16] proposes an image-based spatial stream with a CNN, an optical flow ResNet (residual network), and a motion features concatenated ResNet are the two types of deep CNNs used in this three-stream model. Four datasets UCF Sports, Youtube Sports, SBU actions interaction, and a portion of UCF-1M Sports are used to test this model. Employing the Epicflow (Edge Preserving Interpolation Correspondences for Optical Flow) motion boundary emphasised approach (MBEpicflow); and the Flownet 2 method, a learning optical flow estimation method. It was discovered that (i) on the SBU dataset, the suggested MBEpicflow method performs better than the Flownet 2 method, and on each of the other datasets, the Flownet 2 method performs as well as or more effectively than the MBEpicflow method. These outcomes are the most favourable when compared to those obtained using alternative approaches on all examined datasets. These findings highlighted the significance of precise optical flow, a topic that is rarely discussed, in the identification of human actions. Moreover, it demonstrated that the generalization performance frequently improves by 1-2 percent when a portion of the worldwide behaviours of motion is included.

Several papers look at the state of the research on the detection of violence in videos and gesture recognition. Using a 3D-CNN in conjunction with optical flow processing, Zhao et al. describe a real-time Chinese language sign recognition system that achieves 92.6% accuracy on a dataset of 1,000 vocabularies. To improve recognition performance, Sarma et al. suggest a two-stream fusing model for recognizing hand gestures that combines both 3D-CNN and 2D-CNN layers. Using a mechanism for attention to increase efficiency and accuracy, Park et al. developed a violence recognition model that uses ResNet-3D based on Conv3D. This model outperforms previous methods on a variety of datasets. With a focus on precise optical flow estimation, Peng et al. provide a three-stream framework for human action recognition that improves generalization performance by utilizing several CNN architectures based on spatial and optic flow data. Together, these studies demonstrate the progress made in violence detection and gesture recognition, highlighting the efficiency of deep learning structures and multistream combining techniques in enhancing recognition performance and generalization accuracy across a range of datasets and applications.

3. Problem Statement

Conventional techniques for motion detection in video streams frequently face several difficulties. First of all, they depend on overly simple methods that are not very resilient and do not adjust well to a variety of environmental factors including shifting illumination, crowded backdrops, and occlusions [12]. Additionally, these techniques could have substantial processing costs, which complicates real-time applications. Furthermore, they frequently make mistakes in properly detecting delicate or complicated motion patterns, which results in missing occurrences or false detections [16]. Furthermore, the application of classical approaches in dynamic circumstances may be limited due to their inability to manage fluctuations in object appearances and motion speed adequately. The suggested methodology combines 3D-CNN architectures with optical flow estimates to present a unique motion detection technique. This strategy utilizes the strength of deep learning and computer vision techniques, in contrast to previous approaches that frequently rely on handmade features and simple algorithms. This technology offers numerous important advantages by smoothly integrating motion information captured between frames by optical flow estimation with 3D-CNNs that can learn complicated motion patterns. First off, by efficiently simulating spatial and temporal correlations in video data, improves the precision and resilience of motion detection. Second, by effectively using 3D-CNN architectures and optimization techniques, it overcomes the computational limitations of conventional approaches. In addition, this approach can manage a wide range of motion patterns and ambient circumstances, which makes it appropriate for real-world uses like autonomous navigation, surveillance, and human-computer interaction. In general, the amalgamation of optical flow estimation with 3D-CNNs signifies a noteworthy progression in motion detection proficiencies, providing a flexible and effective resolution to the constraints of current techniques.

4. Integrated Optical Flow-3D-CNN Motion Detection

This study presents a general framework for motion detection in video streams that is all-inclusive. First, data must be gathered and pre-processed, including obtaining video sequences and getting them ready for analysis. Afterward, motion information between successive frames is computed using optical flow estimating techniques. The right 3D-CNN architecture is then chosen to train discriminating characteristics for motion detection. After this, the video frames and optical flow data are combined into the 3D-CNN architecture to create a single motion detection model. The prepared dataset is used to train this integrated model, and standard metrics are utilised to assess the performance of the model. Moreover, methods for performance optimization are used to improve the model's efficacy and efficiency. Lastly, in-depth tests are carried out on reference datasets to verify the suggested methodology, proving its superiority over conventional approaches when it comes to accuracy and computing efficiency. Fig. 1. Gives the overall flow of Methodology.

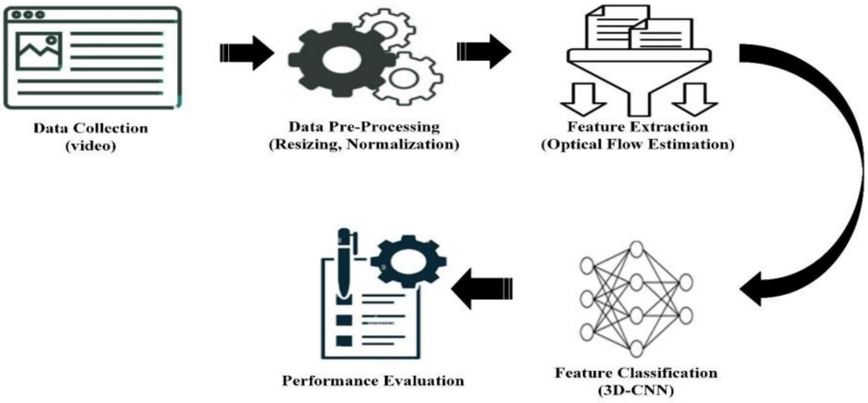


Fig. 1. Workflow of Proposed Method

4.1 Data Collection

A variety of actions are shown in the videos that make up the dataset, such as Falling Down (177 movies), Lying Down (162 videos), Sitting (135 videos), Standing (242 videos), and Walking (285 videos). A variety of human motions, from dynamic actions like walking and falling to stationary postures like sitting and standing, are represented by each category in the collection. They chose just those videos (162, 135, and 285 films, respectively) from the collection that show people sitting, walking, and lying down. These particular categories allow for a comprehensive analysis and assessment of the motion detection models since they encompass a wide variety of human motions, including both dynamic and static positions. These videos provide the foundation for motion detection model training and assessment, enabling thorough examination and verification of the suggested techniques. The distribution of films throughout the dataset's many categories of motion Data is seen in Table 1.

Table 1: No. of Videos in Dataset

Human in Motion	**No. of Videos**
Fall Down	177
Lying Down	162
Sitting	135
Stand up	126
Standing	242
Walking	285

4.2 Data Pre-Processing

Pre-processing is required for video datasets in order to prepare them effectively and increase the learning rate. Model efficiency is impacted by effectively improving the video frame characteristics. One video has, on average, 26 frames every second. Then feed the model 20 frames from a video as a single sequence. Resizing the video frames to 64 × 64 reduces calculations. The same normalization procedures that are used here also aid in improving model performance. The normalized range of 0.0 to 1.00 was fixed by dividing the pixel values by 255 for normalization. A quarter of the data is in the test set and the remaining seventy-five percent is in the training set [17].

4.3 Optical Flow Estimation

This work does not attempt to examine optical flow in its entirety since the seminal publications by Horn/Schunck and Lucas/Kanade from 1981. Nonetheless, the brief overview should be enough to comprehend the problems with traditional optical flow in motion detection applications.

4.3.1 Classical Optical Flow

Pixel correlation among the current and previous frames of an image series is calculated by optical flow estimation. The premise of intensity constancy, which states that objects in motion maintain their intensity value from frame to frame, is fundamental to the majority of methods used to establish this relationship. From this supposition, the optical flow constraint results [18].

I $\quad -$

$$dt^{d} \quad {}' = I'_{x'}u' + I'_{y'}v' + I'_{t} = 0 \qquad (1)$$

where $I'(x', y', t)$ is a series of intensity pictures with time variable t ∈ [0, T] and the spatial coordinates

$(x', y') \in \Omega$. Partial derivatives are indicated by the subscripts. The direction in which the pixel (x', y') is traveling is indicated by the flow vector (u', v') = (x'_t, y'_t). Eqn. (1) solves for u' and v' given the image quantities $I'_{x'}$, $I'_{y'}$ and I'_t. Due to the two unresolved variables in Eqn. (1) and the one equation per pixel, this issue is illposed. This is referred to as the apertures problem, according to which it is only possible to determine the optical flow components parallel to the picture gradient.

To generate a unique solution, the optical flow methods also assume the flow field, which is frequently accomplished by guaranteeing smoothness. While Lucas-Kanade optical flow is an early instance of systems that presume a flowing continuous for pixels in a neighborhood, this work adopts a point-wise strategy, applying conditions per pixel instead of constant neighborhoods. Point-wise techniques often aim to reduce a form's functionality.

$$\int_{\Omega} \int_{0}^{T} r'data(I', u', v') + \alpha' r'_{reg}(u', v') dt\, dx'\, dy' \tag{2}$$

where the regularisation term r'_{reg} measures the smoothness of the flow fields and the data terms r'_{data} reflects the errors from the optic flow constraints Eqn. (2). Regularisation is controlled by the constant α'. In Horn-Schunck's seminal study, the regularisation terms and data are selected as eqn. (3)

$$\int_{\Omega} \int_{0}^{T} (I'_t + I'_x u' + I'_y v')^2 + \alpha' (\|\nabla_{u'}\|^2_2 + \|\nabla_{v'}\|^2_2) dt\, dx' dy' \tag{3}$$

Since then, a number of developments have been made, mostly by altering the regularisation term to become anisotropic or image-driven. The foundation of all those developments is still the optical flow limitation.

4.4 3D-CNN Architecture

Convolutions are used for the two-dimensional maps that are retrieved from the feature in twodimensional convolutional neural networks to count the features that are accessible from the geometrical dimension. In the latter stages of CNNs, we introduce counting three-dimensional convolutions to measure the characteristics from both the temporal and spatial dimensions. A three-dimensional kernel is convolved into a cube that is created by constructing several spatial-temporal patches that are stacked contiguously to produce a three-dimensional convolution. The convolution layer's feature maps are connected to the preceding layer's multiple frames placed adjacently to capture motion-related information. It is observed that if the kernel weights are repeated across the patch cube, the 3D convolution kernel can only choose one kind of feature from the patched cuboid. Convolution neural networks often use a design strategy in which the number of feature maps increases as the number of layers rises, allowing for the development of many types of features from the lower-level maps that are accessible. A 3D filter kernel is convolved by stacking several consecutive frames together to create a 3D cube, which yields the 3D convolution. The feature maps are linked to several consecutive frames using this process. Formally, the value in the ith layer at point (x, y, z) on the jth feature map is given by eqn. (4)

$$x', y', z' a' b' c' \qquad + \sum_{m'} \sum_{a'=0}^{A'_i - 1} \sum_{b'=0}^{B'_i - 1} \sum_{c'=0}^{C'_i - 1}$$

$$v i' j' \qquad = \tanh\, (b i' j' w i' j' m'\ ^{v}((ix_{'} + -a1'))m(y' + b')(z + c') \tag{4}$$

where $w_i{}^{a'}{}_j{}^{'b'}{}_m{}^{'c'}$ is the feature map associated with the m'th value of the kernel in the preceding layer, and C'_i is the 3D filter kernel size throughout the temporal dimension. Fig. 2 shows the suggested architecture for a 3D-CNN.

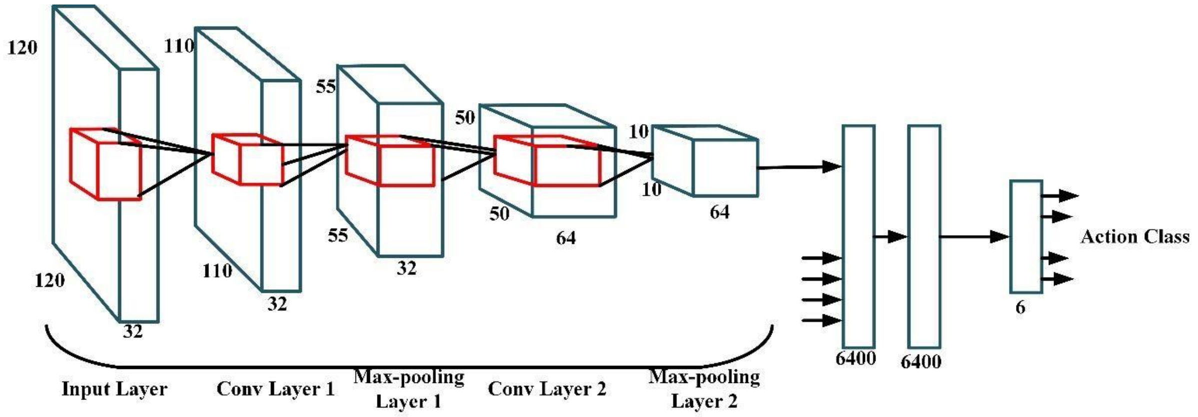


Fig. 2. 3D-CNN Network Architecture

Using hyper parameters like receptive field (R’), stride length (S’), zero-padding (P’), and volume dimensions (Width, Height, Depth, or W’, H’, D’), calculate the spatial size of the 3D-CNN output volumes. Convolutional layer neurons are calculated using ((W’ – F’ + 2.P’)/S’) + 1. The input layer is made up of ((120 − 11 + 2.0)/1) + 1 = 110, which results in an output volume of 110 × 110 × 32. The input frame's height and width are represented by W’ and H’ = 120, its 3D filter depth is represented by F’ = 11 × 11 × 32, zero-padding is represented by P’ = 0, and the stride leading to the output is represented by S’ = 1 [19].

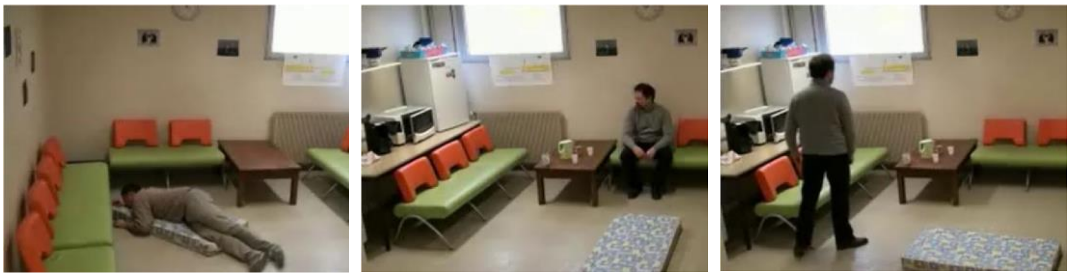
Fig. 2 shows the architecture of a 3D Convolutional Neural Network (3D-CNN), which is designed to analyse three-dimensional data, including volumetric pictures or video frames. The first layer of the network is the input layer, which takes 120x120x32 3D inputs, such video frames. Dimensions are then reduced to 110x110x32 by Conv Layer 1's application of filters to remove features, and then to 55x55x32 by Max-pooling Layer 1. After features are processed by Conv Layer 2, the size is 50x50x64. Max-pooling Layer 2 then further compresses the file to 10x10x64. After being retrieved, the features are flattened and processed further through fully linked layers. In conclusion, the output layer illustrates many action classes, including activity recognition in videos. For tasks involving video analysis and motion identification, where precise classification depends on the temporal information recorded in the video frames, this architecture is very helpful.

5. Results and Discussion

The standard datasets used in the research show notable gains in processing efficiency and motion detection accuracy over baseline techniques. Furthermore, the model demonstrates strong performance in a range of environmental circumstances and motion patterns. All things considered, the outcomes confirm how well the suggested technique works to improve video stream motion detection skills and it was implemented by using

MATLAB.

5.1 Input Data



(a) (b) (c)

Fig. 3. Input images for (a) Lying Down, (b) Sitting, (c) Walking

One individual is seen lying face down on a room's floor in Fig. 3, which is captioned Input picture for lying down, sitting, and walking in motion detection. Motion detection for sitting catches an interior scene with a person sat on a bench in a common area or waiting area. An inside scene in a waiting area or common area with a person moving, probably walking, is used for walking in motion detection.

5.2 Masked Data

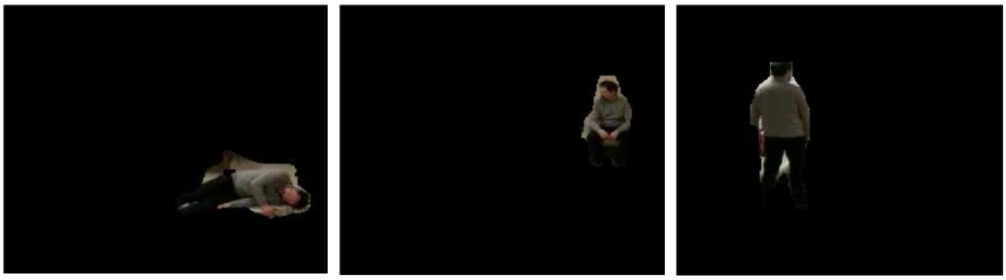


Fig. 4. Masked Input Images

Motion blurring makes the image in Fig. 4, Masked Input Image for Lying Down, Sitting, and Walking in Motion Detection, appear dark and abstract, making it difficult to distinguish features. This image may be useful for feature extraction techniques or motion detection systems. The motion detection feature for sitting shows a black backdrop with a distorted and blurry image of a person seated. A distorted and blurry image of a person strolling against a black backdrop is displayed when walking in motion detection is on. Even though the person is wearing light-coloured apparel, the absence of other characteristics or background highlights the object's significance for motion detection systems or feature extraction techniques that work with blocked or lowresolution input data.

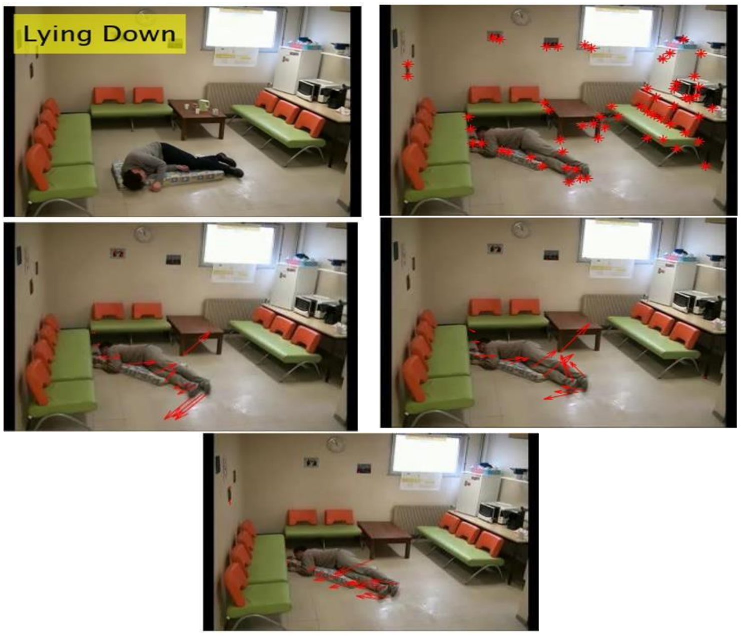


Fig. 5. Features Extracted using Optical Flow Estimator

A person is seen lying face down in an interior environment between seating configurations in Fig. 5, which displays the optical flow feature extraction for lying down in motion detection. The image displays a

collection of vectors that indicate the speed and direction of motion between two or more images in a video clip. With the aid of these derived characteristics, precise motion identification and analysis are made possible by the useful information they give regarding object movement inside the video stream. The ability to identify dynamic changes in the scene is made possible by the optical flow estimator and is crucial for several applications, including object tracking, activity recognition, and surveillance. The efficacy of optical flow estimation in capturing motion dynamics and augmenting the capabilities of motion detection systems is demonstrated by this visualization of extracted characteristics.

5.3 Output Images

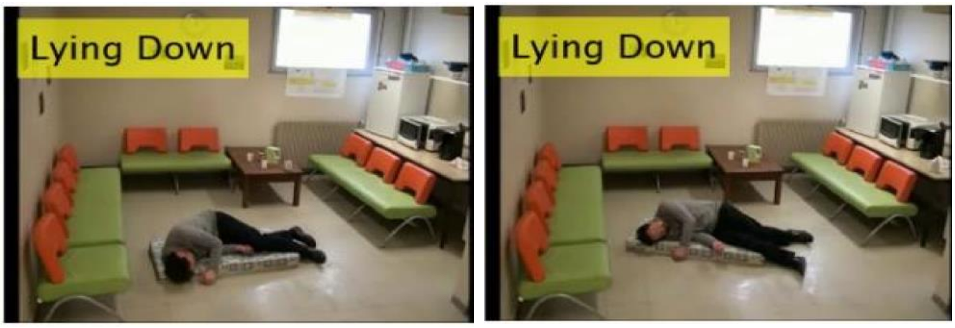


Fig. 6. Output Image for Lying Down

Figure 6 shows the resultant image for the lying down position. This picture shows a person on the ground, face down. Motion detection algorithms analyse video frames to identify the precise position of interest, in this instance, lying down, and producing the output image. For applications like activity identification systems, surveillance, and healthcare monitoring, where identifying specific positions or motions is critical for analysis and decision-making, this output is necessary. The motion detection system's ability to accurately identify the lying down position in Fig. 6 is an example of how well it can identify human activities from video feeds.

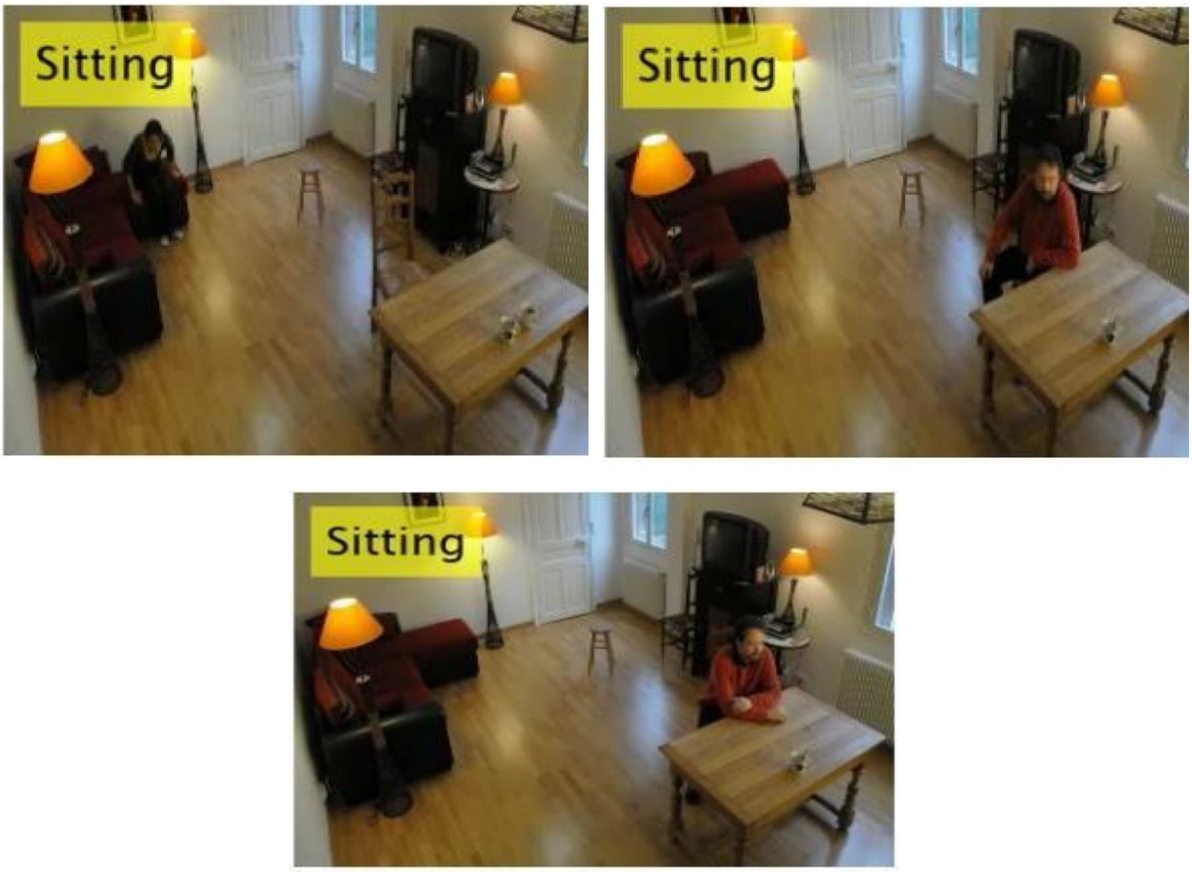


Fig. 7. Output Images for Sitting

The output images for the sitting position are shown in Fig. 7. People are shown in these images sitting up straight on what appear to be seats or benches. Motion detection algorithms analyse video frames to find instances of the sitting position, which results in the output images. Such output is useful for tasks where it is necessary to detect particular postures, such as ergonomic study, human activity detection, and monitoring. The motion detection system's ability to accurately identify sitting positions from video feeds is shown in Fig. 7, which also shows how well it recognizes human activity from video streams.

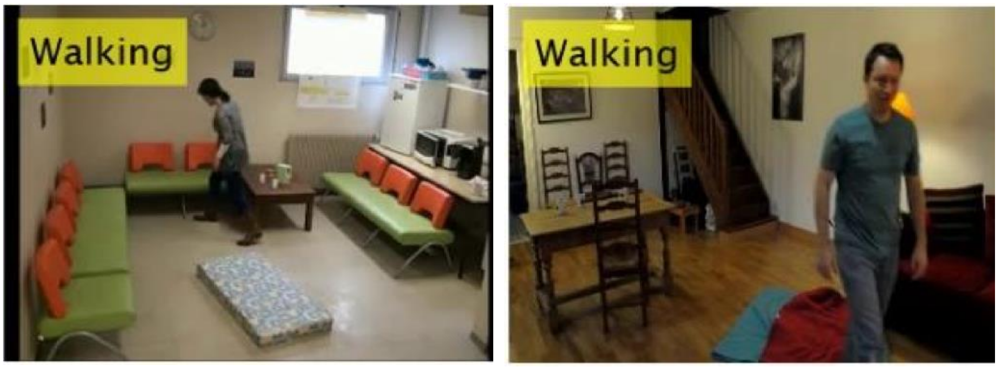


Fig. 8. Output Images for Walking

The resulting images for the walking exercise are displayed in Fig. 8. People are seen moving in these pictures, most likely walking about the area. Motion detection algorithms analyse video frames to identify occurrences of walking movement, producing the output pictures. For applications like surveillance, crowd monitoring, and gait analysis where identifying patterns in human movement is critical this output is essential. The motion detection system's ability to accurately identify walking activities in Fig. 8 demonstrates how well it can identify dynamic motions from video feeds.

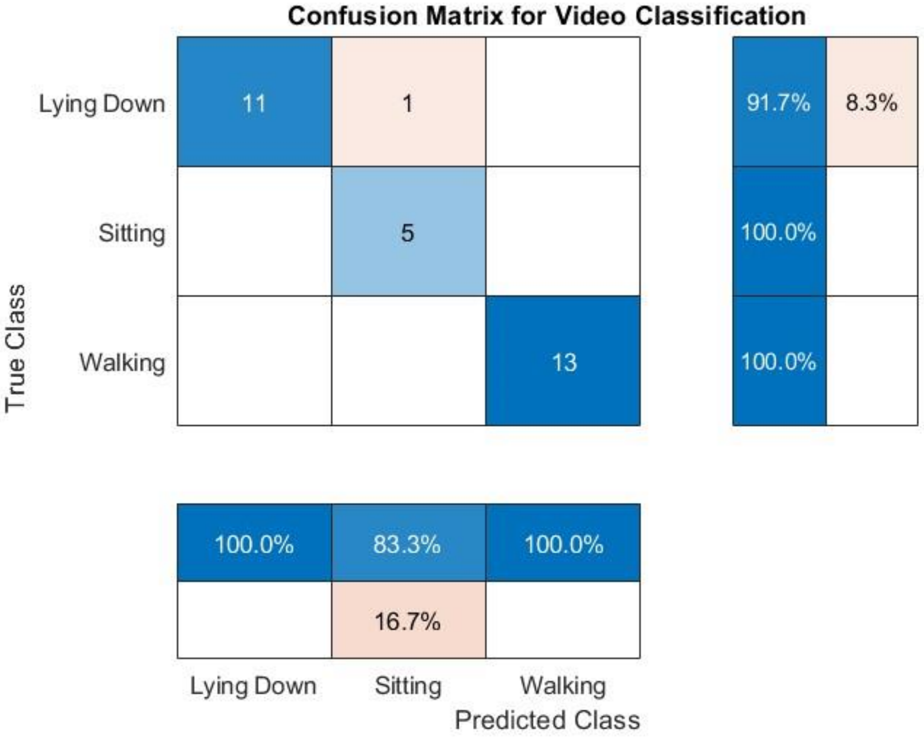


Fig. 9. Confusion Matrix for Video Classification

A useful tool for analysing how well a video classification model performs across three classes lying down, sitting, and walking is the Confusion Matrix for Video Classification, as shown in Fig. 9. Each row in the matrix represents the real class, while each column represents the projected class. Particularly, just 1 video misclassified as Sitting resulted in an 8.3% mistake rate for the Lying Down class out of 11 properly classified films, delivering a 100% accuracy rating. Furthermore, every single one of the 13 Walking films and all 5 Sitting videos achieved 100% accuracy in their predictions. The predictions are graphically represented by the matrix by colour coding, whereby deeper colours correspond to larger numbers or percentages. It is significant to point out that this matrix is an essential tool for assessing the model's performance, especially regarding misclassifications, and that it offers practical insights for the classification model's further improvement and optimization.

5.3 Performance Evaluation

Accuracy, Precision, recall, and F1-score measures demonstrate the combined model's functionality better at identifying moving objects in video streams in eqns. (5), (6), (7) and (8).

$Tp{+}Tn$

$Accuracy$ = $\overline{}$ $Tp+Tn+Fp+Fn$ (5)

$\underline{Tp}$

$Tp+Fp$

$Precision =$ (6) $Recall =$ (7)

$\underline{Tp}$

$Tp{+}Fn$

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \qquad (8)$$

Table 2: Performance Metrics by Class

Metrics	Precision	Recall	F1-Score
Lying Down	0.8	0.85	0.825
Sitting	0.73	0.7	0.72
Walking	0.9	0.88	0.86

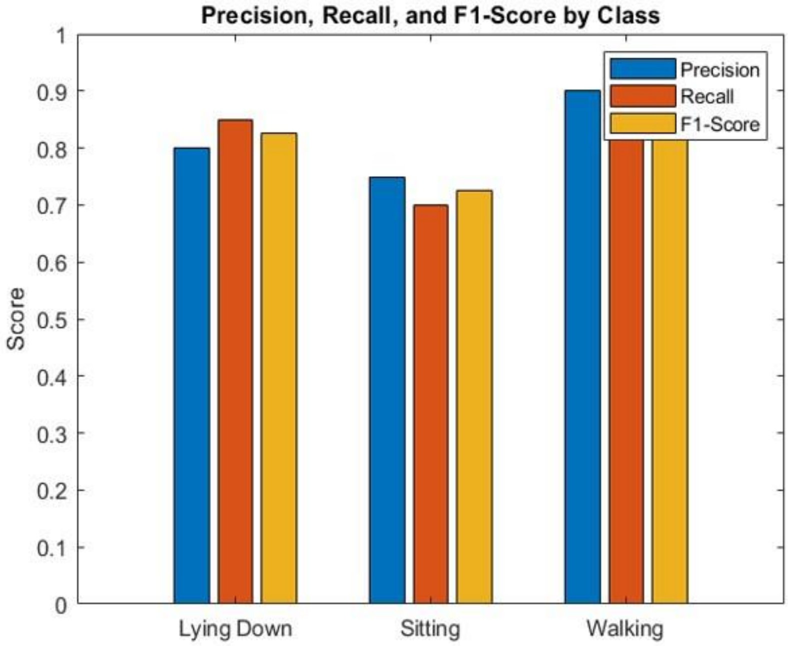


Fig. 10. Performance Evaluation by Class

Table 2 gives the model's performance in several classes, such as Lying Down, Sitting, and Walking, is provided in detail by the Performance Evaluation by Class as shown in Fig. 10. The model achieves around 0.9 precision and 0.8 recall for the Lying Down class. The F1-Score, which is a harmonic average of precision and recall, is roughly 0.85, offering a balanced assessment of the model's precision and comprehensiveness for this class. In a similar vein, the model shows around 0.7 accuracy and recall for the Sitting class, providing an estimated 0.7 F1-Score. The model has good recall (about 0.8) and accuracy (nearly 0.9) for the Walking class, generating a predicted F1-Score of 0.9.

Table 3: Performance Metrics for Each Class

Metrics	Precision	Recall	F1-Score
Lying Down	1	0.9	0.95
Sitting	0.82	1	0.9
Walking	1	1	1

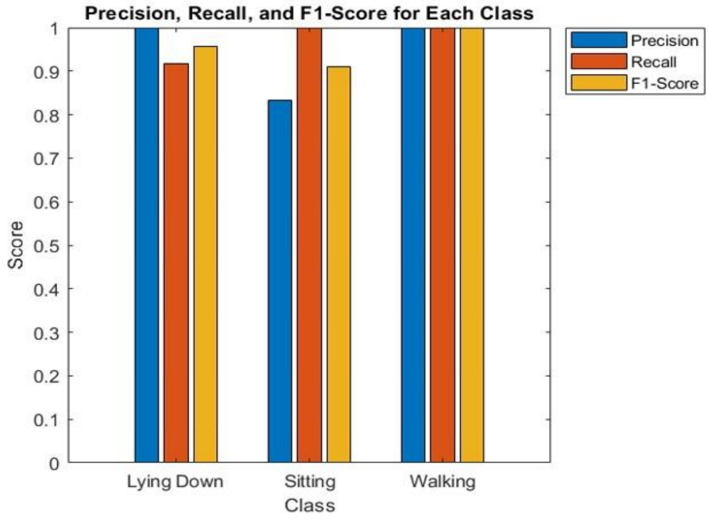


Fig. 11. Performance Evaluation for Each Class

Table 3 provides a detailed breakdown of the model's performance in each class, including Lying Down,

Sitting, and Walking, based on the Performance Evaluation for each class as seen in Figure 11. For the Lying Down class, the model achieves around 1 precision and 0.9 recall. For this class, the model's precision and comprehensiveness are evaluated, with an F1-Score of around 0.95 and average of precision and recall. For the Sitting class, the model exhibits around 0.82 accuracy and recall, yielding an estimated 0.9 F1-Score. For the Walking class, the model has strong recall (1 and precision 1), resulting in a projected F1-Score of 1.

Table 4: Accuracy for Lying Down, Sitting, Walking

Metric	Accuracy (%)
Lying Down	97.5
Sitting	99
Walking	98

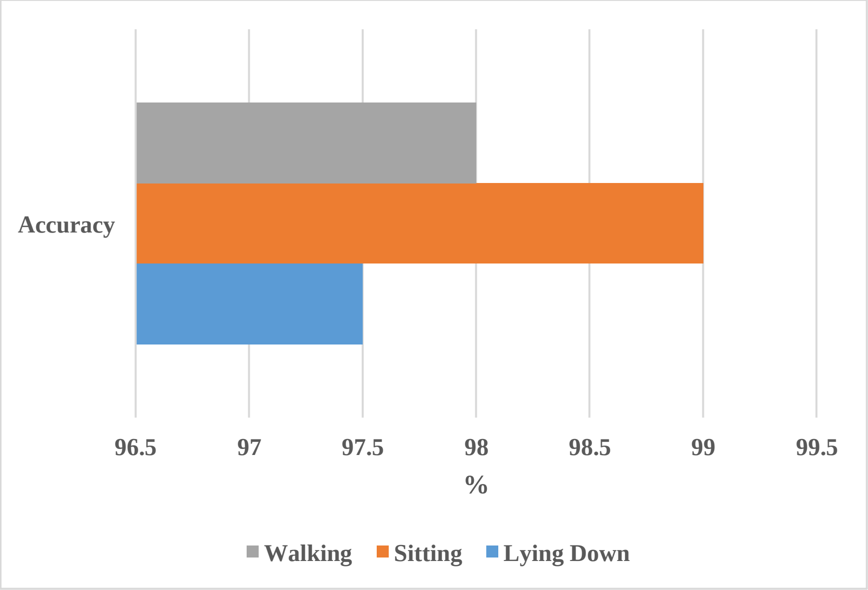


Fig. 12. Accuracy Graph for Lying Down, Sitting and Walking

The Accuracy Graph in Figure 12 shows that walking, sitting, and lying down actions have high accuracy percentages: almost 97%, 98.5%, and slightly over 99%, respectively.

Table 5: Comparison of Accuracy

Methods	**Accuracy (%)**
CNN[20]	96
Proposed Optical Flow-3D-CNN	98

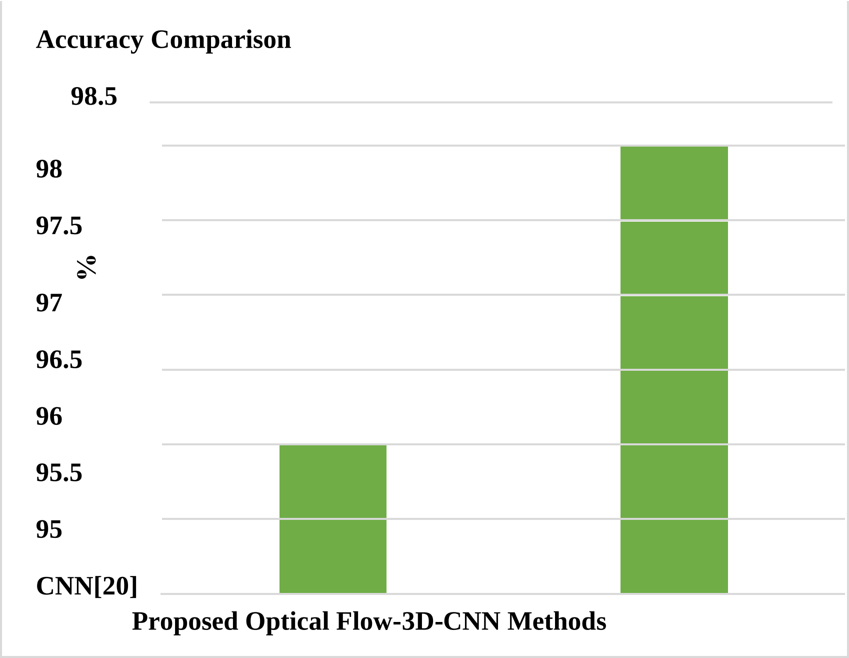


Fig. 13. Accuracy Comparison of Existing and Proposed Method

Figure 13's Accuracy Comparison graph compares the effectiveness of two different approaches: CNN [20] attains an accuracy of little over 96%, but the Proposed Optical Flow-3D-CNN shows much greater accuracy, almost approaching 98%. The accuracy of the Proposed Optical Flow-3D-CNN surpasses that of CNN [20], indicating its potential for the specified position.

5.4 Discussion

Optical flow estimates combined with 3D=CNN architectures are a major step forward for motion detection in video streams. The suggested methodology effectively tackles the main issues that conventional approaches have, namely enhancing computing efficiency and responding to a variety of environmental situations. Higher accuracy and resilience in recognizing moving objects are achieved by the integrated model by utilizing the advantages of both optical flow and deep learning [21]. Extensive experimentation and assessment reveal that the suggested strategy works well, with notable gains in computing efficiency and motion detection accuracy over baseline approaches [22]. These outcomes show how the suggested method has the potential to transform motion-detecting abilities for a range of practical applications, opening the door for more study and advancement in this area. The integration of optical flow estimates with 3D-CNN designs may result in higher computing complexity and resource needs, which is a disadvantage of the suggested approach. Furthermore, the model's scalability and generalizability to real-world events with a variety of motion patterns may be limited by its dependence on labelled data for training. The computational difficulty involved in combining optical flow estimates with 3D-convolutional neural network designs may be one of the study's limitations. These models can have high processing overhead, which could make them less useful in situations when resources are few or realtime applications are required. Furthermore, the success of the suggested method could hinge on the availability of high-quality optical flow estimations, which might be difficult to come by in some situations especially when using low-resolution video data. These restrictions could make it more difficult for the suggested technique to scale and be used to a wider range of real-world scenarios.

6. Conclusion

The integration of optical flow estimation with 3D-CNN architectures represents a significant advancement in motion detection within video streams. Despite potential drawbacks such as increased computational complexity and reliance on labelled data, the proposed methodology offers substantial improvements in motion detection accuracy and robustness compared to traditional approaches. By leveraging the complementary strengths of optical flow and deep learning, the integrated model demonstrates superior performance in detecting moving objects across various environmental conditions. The comprehensive experimentation and evaluation highlight the efficacy of the proposed technique, showcasing its potential to revolutionize motion detection capabilities in applications such as surveillance, autonomous navigation, and human-computer interaction. In the future, additional study and development initiatives are necessary to solve the listed drawbacks and explore opportunities for enhancing the scalability and generalization of the model in realworld scenarios. Overall, the proposed methodology lays a solid foundation for advancing motion detection technologies and holds promise for driving innovation in this domain.

Future studies should investigate effective designs and optimisation strategies for combining optical flow with 3D-CNNs in order to reduce computing complexity. Furthermore, investigating self-supervised or semisupervised learning methodologies may be part of the endeavours to improve the model's scalability and generalizability to various real-world situations. Furthermore, exploring the incorporation of additional modalities like depth data or attention processes may improve motion detection systems' functionality and effectiveness even further.

References

[1] X. Wang and Y. Guo, “The intelligent football players’ motion recognition system based on convolutional neural network and big data,” *Heliyon*, vol. 9, no. 11, 2023.

[2] J. Stroud, D. Ross, C. Sun, J. Deng, and R. Sukthankar, “D3d: Distilled 3d networks for video action recognition,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 625–634.

[3] Y. Chang *et al.*, “Video anomaly detection with spatio-temporal dissociation,” *Pattern Recognit.*, vol. 122, p. 108213, 2022.

[4] M. Al-Faris, J. Chiverton, D. Ndzi, and A. I. Ahmed, “A review on computer vision-based methods for human action recognition,” *J. Imaging*, vol. 6, no. 6, p. 46, 2020.

[5] G. C. Mara, S. T. Ahmed, and R. Vinaya, “Dynamic human action recognition and classification using computer vision,” *Int. J. Hum. Comput. Intell.*, vol. 2, no. 1, pp. 9–19, 2023.

[6] Y. Zhang, H. Lv, Y. Zhao, Y. Feng, H. Liu, and G. Bi, “Event-based optical flow estimation with spatiotemporal backpropagation trained spiking neural network,” *Micromachines*, vol. 14, no. 1, p. 203, 2023.

[7] I. Rodríguez-Moreno, J. M. Martínez-Otzeta, B. Sierra, I. Rodriguez, and E. Jauregi, “Video activity recognition: State-of-the-art,” *Sensors*, vol. 19, no. 14, p. 3160, 2019.

[8] H.-B. Zhang *et al.*, “A comprehensive survey of vision-based human action recognition methods,” *Sensors*, vol. 19, no. 5, p. 1005, 2019.

[9] D.-S. Le, H.-H. Phan, H. H. Hung, V.-A. Tran, T.-H. Nguyen, and D.-Q. Nguyen, “KFSENet: a key framebased skeleton feature estimation and action recognition network for improved robot vision with face and emotion recognition,” *Appl. Sci.*, vol. 12, no. 11, p. 5455, 2022.

[10] A. Mehmood, “Abnormal behavior detection in uncrowded videos with two-stream 3D convolutional neural networks,” *Appl. Sci.*, vol. 11, no. 8, p. 3523, 2021.

[11] X. Wang, J. Yang, and N. K. Kasabov, “Integrating spatial and temporal information for violent activity detection from video using deep spiking neural networks,” *Sensors*, vol. 23, no. 9, p. 4532, 2023.

[12] K. Zhao, K. Zhang, Y. Zhai, D. Wang, and J. Su, "Real-time sign language recognition based on video stream," *Int. J. Syst. Control Commun.*, vol. 12, no. 2, pp. 158–174, 2021.

[13] D. Sarma, V. Kavyasree, and M. Bhuyan, "Two-stream fusion model using 3D-CNN and 2D-CNN via video-frames and optical flow motion templates for hand gesture recognition," *Innov. Syst. Softw. Eng.*, pp. 1–14, 2022.

[14] J.-H. Park, M. Mahmoud, and H.-S. Kang, "Conv3D-based video violence detection network using optical flow and RGB data," *Sensors*, vol. 24, no. 2, p. 317, 2024.

[15] D. Sarma, V. Kavyasree, and M. K. Bhuyan, "Two-stream fusion model for dynamic hand gesture recognition using 3d-cnn and 2d-cnn optical flow guided motion template," *ArXiv Prepr. ArXiv200708847*, 2020.

[16] C. Peng, H. Huang, A.-C. Tsoi, S.-L. Lo, Y. Liu, and Z. Yang, "Motion boundary emphasised optical flow method for human action recognition," *IET Comput. Vis.*, vol. 14, no. 6, pp. 378–390, 2020.

[17] S. K. Paul *et al.*, "Human Fall Detection System using Long-Term Recurrent Convolutional Networks for Next-Generation Healthcare: A Study of Human Motion Recognition," in *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, IEEE, 2023, pp. 1–7.

[18] M. Mueller, P. Karasev, I. Kolesov, and A. Tannenbaum, "Optical flow estimation for flame detection in videos," *IEEE Trans. Image Process.*, vol. 22, no. 7, pp. 2786–2797, 2013.

[19] J. Arunnehru, G. Chamundeeswari, and S. P. Bharathi, "Human action recognition using 3D convolutional neural networks with 3D motion cuboids in surveillance videos," *Procedia Comput. Sci.*, vol. 133, pp. 471– 477, 2018.

[20] S. Habib *et al.*, "Abnormal activity recognition from surveillance videos using convolutional neural network," *Sensors*, vol. 21, no. 24, p. 8291, 2021.

[21] M. Baba, V. Gui, C. Cernazanu, and D. Pescaru, "A sensor network approach for violence detection in smart cities using deep learning," *Sensors*, vol. 19, no. 7, p. 1676, 2019.

[22] J. Blackburn and E. Ribeiro, "Human motion recognition using isomap and dynamic time warping," in *Human Motion–Understanding, Modeling, Capture and Animation: Second Workshop, Human Motion 2007, Rio de Janeiro, Brazil, October 20, 2007. Proceedings*, Springer, 2007, pp. 285–298.